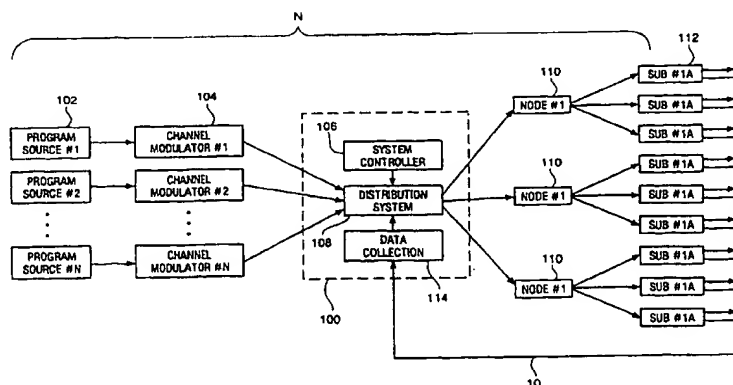




INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>6</sup> : <b>H04N 7/173</b>		<b>A1</b>	(11) International Publication Number: <b>WO 98/37696</b>
			(43) International Publication Date: 27 August 1998 (27.08.98)
(21) International Application Number: PCT/US98/03181		(81) Designated States: AU, CA, CN, JP, MX, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(22) International Filing Date: 17 February 1998 (17.02.98)			
(30) Priority Data: 60/037,354 21 February 1997 (21.02.97) US		Published With international search report.	
(71)(72) Applicant and Inventor: HERZ, Frederick, S., M. [US/US]; Box 625, Canaan Valley, Davis, WV 26260 (US).			
(72) Inventors: SMITH, Jonathan, M.; 771 Princeton-Kingston Road, Princeton, NJ 08540-4165 (US). WACHOB, David; 8379 Glen Road, Elkins Park, PA 19117 (US).			
(74) Agents: DUFFT, Donald, M. et al.; Duft, Graziano & Forest, P.C., Suite 140, 1790 30th Street, Boulder, CO 80301-1018 (US).			

(54) Title: BROADCAST DATA DISTRIBUTION SYSTEM WITH ASYMMETRIC UPLINK/DOWNLINK BANDWIDTHS



(57) Abstract

The broadcast data distribution system distributes directory and indexing information for the selection of viewing choices in broadcast and multicast networks with asymmetric uplink/downlink bandwidths. Examples of such systems include Digital Broadcast Satellite (DBS) and Cable Television (CATV) systems. More particularly, a preferred embodiment of the invention co-locates shared directory resources at directory servers which can provide the benefit of technology improvement through storage multiplexing to directory clients, which are assumed to have smaller, more heterogeneous and slowly-evolving memory and processing resources. The directory information stored at the directory clients is a subset of the information maintained at the directory servers selected on the basis of subscriber interest. This system automatically constructs both a target profile for each target object (program) that is broadcast, as well as a "target profile interest summary" for each subscriber, which target profile interest summary describes the subscriber's interest level in various types of target objects. The system then evaluates the target profiles against the subscriber's target profile interest summaries to generate a subscriber-customized rank ordered listing of target objects most likely to be of interest to each subscriber, so that the subscriber can select from among these potentially relevant target objects, which were automatically selected by this system from the plethora of target objects available on the data distribution system.

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

## BROADCAST DATA DISTRIBUTION SYSTEM WITH ASYMMETRIC UPLINK/DOWNLINK BANDWIDTHS

### FIELD OF INVENTION

5           This invention relates to the field of data distribution systems and comprises a method, extant in the data distribution system, to enhance the efficiency of the data transfer to thereby reduce the memory requirements of a subscriber network terminal. This enables the subscriber network terminal to be implemented as a simple, low cost terminal device which can interface with an ever changing and  
10   expanding capacity data distribution system and yet provide the subscriber with the information required to efficiently access the plethora of data available to the subscriber via the data distribution system.

### PROBLEM

          It is a problem in the field of data distribution systems to provide a subscriber  
15   with a simple, low cost terminal device which can interface with a changing data distribution system and yet provide the subscriber with the information required to enable the subscriber to efficiently access the plethora of data available to the subscriber via the data distribution system. The data distribution systems of interest are typically cable television (CATV) or Digital Broadcast Satellite (DBS)  
20   systems, which concurrently transmit multimedia data streams on a plurality of channels to a subscriber terminal adapter, which has limited memory and functionality, and which functions to receive a selected one of the plurality of broadcast channels for display on a display device, such as the subscriber's television set. In addition, the Internet has the inherent capability to be included in  
25   this class of data distribution systems due to its ability to broadcast data to a distributed clientele.

          In the 1960s and 1970s, data transmission capabilities were limited and therefore few broadcast channel selections were available to a subscriber of CATV or DBS networks. Today, typical DBS and CATV networks provide several tens of  
30   channels, and it is expected that in the near future hundreds of channels will be concurrently broadcast to the subscribers via CATV and DBS networks. This is desirable from both subscriber and content-provider perspectives, as increasingly specialized information can be generated, partitioned and delivered to selected sets of subscribers, whose interests match the content of the broadcast information. An

unfortunate by-product of this plethora of information is information overload, and a great deal of a subscriber's time is wasted "channel surfing." This channel surfing comprises the subscriber attempting to locate relevant and/or interesting information from the wealth of broadcast information that is available by methodically switching among the available broadcast channels. This process is inefficient and represents a substantial reduction of the usable information delivered to the subscriber due to the subscriber's inability to locate the desired information in a timely manner.

A problem with the delivery of information to subscribers is the implementation of the subscriber's network interface device which enables the subscriber to interface with the data distribution system. Several scenarios exist for the subscriber network interface device. One is the use of intelligent, high functionality devices such as personal computers. This is presently an unlikely scenario for three reasons. First, personal computers are present in less than 30% of US households, while CATV is presently available in 65% of households. DBS has a lower penetration than personal computers or CATV systems (presently less than 5%) but is expected to increase over time. The problem with requiring the purchase of a relatively expensive subscriber network interface device is that it inhibits unit sales by greatly restricting the market potential. Second, the personal computer market has been characterized by component providers competing to sell technologies, such as network adapters and monitors. In addition, many personal computers are pieced together from components which are manufactured by disparate vendors and the resultant product is sold to consumers with a particular price/performance characteristic. This creates an almost impossible environment from the view of engineering a subscriber network interface device with widespread interoperability, which is necessary from a business point of view for CATV fare. Different software versions and their incompatible requirements further complicate interoperability. Third, the use of a personal computer further exacerbates the risk of the "flashing 12:00" phenomenon seen with VCRs, since the complexity of the personal computer, when used as the subscriber network interface device, inhibits technologically-limited subscribers from gaining access to the information.

A much more likely scenario is the use of a specialized terminal adapter, presently referred to in the media as a "set-top box." In this scenario, the subscriber network interface device comprises a standard television set which is equipped with a terminal adapter, which is specialized to the task of accessing multimedia streams for the subscriber and presenting this information via the television set. This terminal adapter contains a microprocessor as well as any logic necessary for selecting, decompressing, and demultiplexing the concurrently transmitted multimedia data streams. The terminal adapter also handles out-of-band signaling tasks, such as authorizing access to restricted material and executing any protocol required for subscriber/service-provider interaction. Several competing factors are at work in optimizing the terminal adapter. First, as service costs are borne by the service provider in many cases, and service outages annoy subscribers, the terminal adapter must be robust. Second, the terminal adapter must be flexible enough to accommodate any anticipated information delivery scenario within the limits of the DBS or CATV infrastructures. Finally, the terminal adapter must be low-cost enough to either be within the reach of subscriber purchase or a quick cost recovery through leasing to the subscriber. It is important to note that in this latter scenario, while the cost may be amortized over a considerable period, the monthly bill to the subscriber should not exceed the cost of film rentals or be excessive in the view of government regulators.

The terminal adapter includes both a processor and memory, which represents a major cost factor. This memory supports the storage and processing of data and, for many computing applications, the easiest way to speed up a computation task is to add more memory. Unfortunately, there are several factors which preclude equipping the terminal adapter with substantial memory. First is the cost of the memory, which stems both from the amount of memory required and the type of devices required for the terminal adapter. Non-volatile RAMs are required for at least part of the data storage to handle the reliability consequences of power failures or subscribers disconnecting the terminal adapter. The second factor is the need to avoid updating the hardware of the terminal adapter, since this is an expensive operation to implement. An information distribution architecture which requires extensive memory at the terminal adapter forces technicians to make

updates at subscriber sites on a regular basis as earlier installations became outdated.

New-generation CATV "converter" boxes provide storage for information relative to presently broadcast or upcoming programs (or more generally, multimedia data streams). The storage of this "program guide" information in the terminal adapter is quite memory-intensive, on the order of 170 Kbytes, and with the boom in channel availability, will require increases which are proportional to the number of additional channels. The large memory required in the terminal adapter is a consequence of the lack of information about the viewer's tastes and interests, which information could be used to select subsets of the descriptive information for local storage in the terminal adapter. Since no such information is available, all indexing and descriptive information regarding the multitude of concurrently broadcast programs must be transmitted and retained in the memory of the terminal adapter. Thus, there presently is no effective terminal adapter/data distribution system architecture or operation that addresses the subscriber's need to have program guide information in the terminal adapter and yet is frugal in its use of expensive memory.

A further application of this technology is in the field of data retrieval, wherein the subscriber requests data of a specific nature, rather than standard broadcast program fare. In this instance, the information distribution system must discern what information is being requested by the subscriber and provide an efficient mechanism for delivering this information. The delivery of such information is complicated when the subscriber is mobile and requests information from an information source which is not the subscriber's "home location" information source. Such instances include the cellular communications or personal communication system applications as well as the general Internet access application. The subscriber can access any of these services from any physical location by interconnecting with a local purveyor of such services. The information requested by the subscriber can be locale specific, or may be customized for time of day, and the existing information distribution systems cannot produce such specific requests in a manner that is convenient for the subscriber.

### SOLUTION

The above problems are solved and a technical advance achieved in the art by the data distribution system of the present invention. The underlying architecture of the data distribution system is typically a hierarchical network, such as cable television. The data distribution system therefore provides a hierarchical filtering of the data content of information requested by or available to a subscriber to reduce the effective bandwidth required to service a subscriber's request for information.

To accomplish effective filtering, this data distribution system automatically constructs both a target profile for each target object (program or information response) that is broadcast, as well as a "target profile interest summary" for each subscriber, which target profile interest summary describes the subscriber's interest level in various types of target objects. The data distribution system then evaluates the target profiles against the subscribers' target profile interest summaries to generate a subscriber-customized rank ordered listing of target objects most likely to be of interest to each subscriber, so that the subscriber can select from among these potentially relevant target objects, which were automatically selected by this system from the plethora of target objects available on the data distribution system. The target objects themselves are divided into a plurality of information segments, which enables the data distribution system to provide the subscriber with manageable amounts of data. These content based similarity measurements help to determine which of the information segments to transmit and the order of transmission to the subscriber. The similarity measures can also be used to prefetch information segments and deliver these segments to the subscriber in a bandwidth saving manner. This capability is extensible to mobile subscribers who change locale and may require the delivery of location specific information, as a function of their changing locale.

This architecture offers several advantages. First, technology upgrades can be done in the core of the data distribution network, without requiring access to subscriber terminal equipment. Second, there are significant savings in data storage which results from multiplexing this directory information across populations of subscribers. Third, even an extremely limited amount of memory in the terminal

adapter can handle complex information streams and subscriber interests with complete generality, due to the segmentation of the target objects.

The data distribution system is comprised of clients, which are the terminal adapters with extremely limited data storage capacity, and servers, which contain data, such as multimedia information (such as live program feeds and/or video-on-demand) and descriptive information (directory information, schedules, indices) for the multimedia information. Clients are assumed to be under subscriber physical control, while servers are embedded in the physical infrastructure of the data distribution system. The keys to a limited data storage architecture are the intelligent use of local storage, optimized through memory management algorithms, the use of subscriber target profile interest summary information, and the reliance on a communications protocol between client and server to update local storage through server updates and client requests. This results in a distributed architecture for directory information which is robust in the face of change, is low-cost, and utilizes the CATV/DBS infrastructure itself to preserve these properties. In practice, the client is the terminal adapter which comprises the "set-top box." The overall algorithm used by the server to provide the directory information for the client is as follows:

1. If new directory information is loaded into the server, the server stores the information in a schedule data structure. If this information represents an update of previously received directory information, the recipients of the previous directory information (the "copy set") are informed that they should delete that previously received directory information. This ensures that while not necessarily complete, the directory information stored at each client is correct.

2. If the server receives a request from a client for directory information, it obtains the most current directory information from its own storage and sends this to the client.

3. The server periodically sends a time-of-day and date indicator to the client system to ensure synchronization between client and server.

In this manner, the data distribution system maintains synchronization with the subscriber terminal adapters and ensures that the directory information stored therein is maintained at present values. The use of subscriber target interest profile



summary filtering of broadcast programs significantly reduces the quantity of directory information required to be stored in the memory of the subscriber terminal adapter, yet functions to direct the subscriber to programs of interest in an efficient manner.

- 5 In the instance of information that is other than broadcast programs, the system uses the target profile interest summary for the subscriber to determine what information is of interest to the subscriber. This data can be used to create similarity measures to thereby implement intelligent precaching of data for transmission to the subscriber. The order as well as content of the data transmitted
- 10 to the requesting subscriber is varied to maximize the efficiency of the data transfer given the limitations of the transmission bandwidth. For mobile subscribers, the system can customize the information delivery as a function of subscriber locale and also precache the data in a locale in anticipation of the arrival of the subscriber.

#### **BRIEF DESCRIPTION OF THE DRAWING**

- 15 Figure 1 illustrates in block diagram form a typical architecture of a multimedia distribution system having a customer preference profile system incorporated therein;

Figure 2 illustrates in flow diagram form the method of creating an agreement matrix;

- 20 Figure 3 illustrates in flow diagram form the method of selecting virtual channels;

Figure 4 illustrates in block diagram form a typical subscriber terminal adapter used in conjunction with the system of Figure 1;

- 25 Figure 5 illustrates in block diagram form the software/firmware architecture of the typical subscriber terminal adapter of Figure 4;

Figure 6 illustrates in flow diagram form the operation of the similarity determination process;

Figure 7 illustrates in flow diagram form the operation of the program matching process;

- 30 Figure 8 illustrates the bandwidth utilization of the data distribution system; and

Figures 9 and 10 illustrate various client-server interconnections.

### DETAILED DESCRIPTION

Figure 1 illustrates in block diagram form a typical architecture of the data distribution system of the present invention. The data distribution system is comprised of clients, which are stored in the terminal adapters 112 with extremely limited data storage capacity that are located in the subscriber's premises, and servers, which are data processing 100, data transmission 108 and data storage elements 110 in the network (which could be located at the headend) which contain both multimedia information (such as live feeds and video-on-demand) and descriptive information (directory information, schedules, indices) for the multimedia information. Clients are assumed to be under subscriber physical control, while servers are embedded in the physical infrastructure of the data distribution system. The basic network structure which interconnects the clients and servers is a high-bandwidth multicast downlink/low-bandwidth unicast uplink network. These bandwidth limitations are a matter of practical implementation, and the inequality of bandwidth in the downlink and uplink paths are not a technological necessity.

The data distribution system may be usefully applied within the context of set-top box like devices such as personal digital assistants or network computers, or personal computers used as a form of set-top box, as a method of reducing response time as observed by users. V. N. Padmanabhan and J. C. Mogul, in an article titled "Using Predictive Prefetching to Improve World Wide Web Latency," ACM SIGCOMM Computer Communication Review, July 1996, disclosed a method of prefetching information stored in the Hypertext Transport Protocol (HTTP) of the Internet-based World Wide Web. In essence, the method used "links" to other documents embedded in an HTTP-format file as hints that those links should be followed in prefetching data; that is, the linked documents should be fetched in anticipation of the user's desire to follow the links to those documents. The data distribution system provides two enhancements to this scheme. First, it provides a technological means by which the prefetched data can be intermixed with on-demand data to provide overall improvements in response time to a large population of HTTP/WWW users, with reduced memory requirements. Second, the data distribution system, which views the downlink as a fixed capacity resource, provides a general scheduling method embodying techniques such as user

preferences to prefetch when slots or bandwidth are underutilized, to preemptively reduce future demand for bandwidth. The general technique of using similarity-informed prefetching is described at length in copending patent application titled "Pseudonymous Server For System For Customized Electronic Identification Of Desirable Objects", U.S. Patent Application Serial No. 08/550,886, filed October 31, 1995.

### **Basic Definitions**

Relevant definitions of terms for the purpose of this description include: (a) an object available for access by the subscriber, which is typically electronic in nature, is termed a "target object", (b) a digitally represented profile indicating that target object's attributes is termed a "target profile", (c) a profile holding a subscriber's attributes is termed a "subscriber profile", (d) a summary of digital profiles of target objects that a subscriber likes and/or dislikes, is termed the "subscriber target profile interest summary" of that subscriber, (e) a collection of target objects with similar profiles, is termed a "cluster," (f) an aggregate profile formed by averaging the attributes of all target objects in a cluster, termed a "cluster profile." In the CATV/DBS examples described herein, the term "program" is used in place of "target object" for clarity of description.

### **Network Elements and System Characteristics**

The various elements interconnected by the data distribution network N as shown in Figure 1 can be divided into two classes and grouped as illustrated in Figure 1: clients and servers. In the CATV example, the plurality of clients are each an individual subscriber's terminal adapter 112 which is connected to servers 102 (program sources) via data communication links. The clients are not continuously on-line, since they typically serve a single subscriber household and the usage pattern is a function of subscriber activity. A server 102 is typically a program source system that is presumed to be continuously on-line and functions to either broadcast a series of programs in a predetermined ordering or retrieve and broadcast a program specifically selected by a subscriber (video on demand). The server 102 is equipped with persistent storage, such as a magnetic tape or magnetic disk data storage medium, and is interconnected with clients via data communications links. The data communications links can be of arbitrary topology

and architecture, and are described herein for the purpose of simplicity as a typical cable television topology. In particular, the present CATV architecture is a wire plant which emulates broadcast television. In all practical aspects, it can be viewed as a tree, from which information is taken out by subscribers at the leaves and put  
5 in by the program source at the root. For example, the installation of a satellite dish at the root of the network, which feeds the remaining infrastructure, enables the output of the satellite dish to be multiplexed across a number of subscribers connected via coaxial cable. A digital broadcast satellite system can also be viewed as a rather flat tree (with the root in orbit) fed from several sources.

10 In the cable television system, a plurality of program sources 102 generate programs which are modulated by a plurality of channel modulators 104 and transmitted to a distribution system 108. The distribution system 108 is part of a head end 100 which retransmits the received programs via respective nodes 110 to subscriber terminal adapters 112. The subscriber terminal adapters 112 contain  
15 software and headend system 100 also contains software extant on system controller 106 to determine which programs are of likely interest for each subscriber. This program selection software controls the programming of virtual channels to the subscriber terminal adapter 112 to select from among plurality of channels that are broadcast over the network N to the head end 100. The  
20 subscriber can use the subscriber terminal adapter 112 to select from the automatically generated virtual channels or the actual broadcast channels available from the head end 100.

Passive feedback relating to subscriber preferences for use in future program scheduling can be incorporated in this data distribution system via a  
25 subscriber data collection element 114 which is connected to the head end 100. In this embodiment, the generation of a target profile interest summary for a subscriber can be implemented at head end 100, which enables the head end operator to ascertain which programming is desired by the collection of subscribers, or it can be implemented in the subscriber terminal adapter 112. In the former  
30 case, the program selection software acts as a filter to enable virtual channel data to be created at the subscriber terminal adapter 112. The return signals to the head end 100 can be carried via the CATV system or via telephone lines 10 using

a wired return path for data collection. Thus, the basic network structure which interconnects the clients and servers is a high-bandwidth multicast downlink/low-bandwidth unicast uplink network.

While the above description is focused on the domain of channel selection in a CATV broadcast television setting, the multiplicity of channels enabled by digital technology and compression technology can allow the use of these channels in a mode very much like a true "Video on Demand" (VoD) system. The multiplicity of channels could be used in a "frequency-hopped" mode (from the perspective of any given stream) to provide access to bursty data such as MPEG-compressed streams of video. The descriptive information could be used for program selection, display, and to customize the selections and scheduling of the program.

#### **Subscriber Terminal Adapter**

Figure 4 illustrates in block diagram form the overall architecture of a typical subscriber terminal adapter for the CATV application, while Figure 5 illustrates the architecture of the software/firmware resident therein. In particular, the program materials and the corresponding program content profile/directory information are received from the head end 100 via tuner 402 to the data receiver 404. Alternatively, the content profile/directory information can be received via a dedicated data channel (illustrated by the dotted line in Figure 4) which can bypass the tuner 402. In either case, the program content profile/directory information that are received can be a subset of the materials available at the head end 100, with the materials representing a distillation of the available information, modulated by the target profile interest summary for this subscriber.

If the program materials transmitted by the head end 100 are scrambled, the scrambled received program materials are transmitted from the tuner 402 to the descrambler 416 prior to being processed by processor 406 and/or modulated by modulator 418 for display. If the tuner 402 selects a one of the concurrently broadcast channels, the content profile/directory information is supplied to processor 406 either directly if the received program contains content profile/directory information in its vertical blanking interval, or via data receiver 404. Processor 406 reviews the content profile/directory information associated with the program material to determine whether there is a correspondence between the

target profile for this program material and the subscriber's target profile interest summary. Input from the subscriber is available via the subscriber terminal adapter 112 generating display menus via on screen menu device 414 and the subscriber using remote control device 408 and infrared receiver 410 as is well known in the  
5 broadcast television art. The subscriber target profile interest summary is stored in memory 412 and is used to ascertain the appropriateness of the various programs broadcast to the subscriber terminal adapter 112. The processor 406 selects a plurality of the received programs as corresponding to the subscriber target profile interest summary and assigns these programs to virtual channels for  
10 consideration by the subscriber. The virtual channels are presented to the subscriber via the television set as is well known. The subscriber can then select one of the virtual channels or any other available channel to view a program selected by the subscriber.

A telephone interface 420 provides a communication path in a unicast mode  
15 from the subscriber terminal adapter 112 to the head end 100 via data collection circuit 114 to enable processor 406 to transmit the collected viewing habit data and subscriber target profile interest summary data from memory 412 to the head end 100. The subscriber target profile interest summary data can optionally be maintained in confidence by the use of encryption device 422 to prevent  
20 unauthorized access to this data. In addition, the subscriber target profile interest summary data can be returned to the head end 100 via the use of RF modulator 424 to enable the bidirectional use of the head end 100-subscriber terminal adapter 112 link.

The software/firmware which is operational in the subscriber terminal adapter  
25 112 to perform the control functions described above is illustrated in block diagram form in Figure 5. In particular, an authorized service segment 502 of memory 412 is used to store data indicative of the programs which the subscriber is authorized to receive as part of the subscription to the CATV network. The content profile/directory information associated with these authorized programs are stored  
30 in directory segment 504, which data is used by processor 406 to calculate the listing of programs which the subscriber is likely to wish to view (termed agreement matrix). The agreement matrix 508 is stored in memory 412 as is an associated

time of day 510, which is either locally generated or received from the head end 100. From the agreement matrix 508, processor 406 calculates a list of virtual channels which is entered into the program guide listing 514. The program guide listing 514 is displayed to the subscriber via the subscriber television set in well known fashion so that the subscriber can select a program from the available choices. Once the subscriber selects a program from the program guide listing 514, the processor 406 activates channel selector 512 to tune to the channel which contains the broadcast of the selected program. The subscriber can also view the generated subscriber target profile interest summary to thereby enable the subscriber to manually adjust or modify the data contained therein to reflect changes in the subscriber's desired viewing habits. The integrity of the subscriber target profile interest summary data is maintained by use of a subscriber identifier 518 which ensures that the subscriber input device 516 is not manipulated by an individual other than the designated subscriber.

#### 15 **Data Storage Architecture**

The keys to a limited data storage architecture in this data distribution system are the intelligent use of local storage (memory 412), optimized through memory management algorithms and the reliance on a communications protocol between client and server to update local data storage (memory 412) through server updates and client requests. This results in a distributed architecture for directory information which is robust in the face of change, is low-cost, and utilizes the CATV infrastructure itself to preserve these properties. Since the client is the terminal adapter 112 in the form of the "set-top box," the overall data distribution algorithm used by the server is as follows:

- 25 1. If new directory information is loaded into the server, the server stores the information in a schedule data structure. If this information represents an update of directory information previously received from the server, the recipients of the previous directory information (the "copy set") are informed that they should delete that previously received directory information and replace it with the newly received  
30 directory information. This ensures that while the directory information at a client may not necessarily be complete, it is correct.

2. If the server receives a request from a client for directory information, it obtains the most current directory information from its own storage and sends this to the client.

3. The server periodically sends a time-of-day and date indicator information to the client system to ensure synchronization between client and server.

In this manner, the data distribution system maintains synchronization with the subscriber terminal adapters 112 and ensures that the directory information stored therein is maintained at present values. The use of subscriber target interest profile summary filtering of broadcast programs significantly reduces the quantity of directory information required to be stored in the memory of the subscriber terminal adapter, yet functions to direct the subscriber to programs of interest in an efficient manner.

#### **Data Transfer Process**

Figure 2 illustrates in flow diagram form the process by which the directory information is provided to the subscriber terminal adapter 112 and the agreement matrix is generated. This process can be implemented in the data distribution system or in the subscriber terminal adapter 112, as a function of the transmission bandwidth available as well as the data storage capacity of the subscriber terminal adapter 112. It is assumed for the purpose of this description that the process takes place in whole or at least in part, in the data distribution system, and in particular the system controller 106 of the head end 108. The schedule of available programs and their characteristics (content profiles) is created and stored in a database in the head end 100 at step 201. The subscriber target profile interest summaries are downloaded for a given time period from the subscriber terminal adapter 112 and, at step 204, stored in a database. The subscriber target profile interest summaries comprise the subscribers' preferences for programs and preferably vary as a function of time of day, which is typically a function of the identity of the multitude of viewers in the subscriber household.

The content profiles received with the electronic program guide data are matched with subscriber-specific preferences (subscriber target profile interest summaries) to create an agreement matrix at step 206. Once the agreement matrix has been created, the programs with the highest values for correspondence to the



subscriber's preferences are prioritized and selected for presentation to the subscriber as virtual channels at step 208. The virtual channel assignment is processed in priority order, with the most desirable programs being assigned first to virtual channels until all desired programs are assigned. However, the virtual  
5 channel assignment is terminated if there are insufficient virtual channels for the number of programs matched by the agreement matrix. It is preferable that the subscriber target profile interest summaries and content profiles be updated to allow for changes in the subscriber preferences as well as to correct for errors in the initial determination. At step 210, the system controller 106 maintains a record  
10 of programs accessed by the subscriber and length of access time. At step 212, the subscriber target profile interest summaries are updated in the subscriber terminal adapter 112.

Once the agreement matrix is created, the selection of virtual channels can be accomplished by means of the process illustrated in flow diagram form in Figure  
15 3. At step 302, determine which subscriber target profile interest summary is active for the time period to be scheduled. The available programs are determined by selecting a subset of all available programs indicative of the extent of the subscriber's program subscription. An agreement matrix is created at step 304 for the available programs and the most popular programs for the subscriber are  
20 selected. At step 306 it is determined whether all of the subscriber's preferences are matched and whether any additional programs are available to match any subscriber preferences that remain. At step 308, the scheduling process is exited if all preferences are matched with the available programs. If not, at steps 310 and 312 the preference list is decremented to eliminate those preferences matched and  
25 processing returns to step 304.

#### **Digital Broadcast Satellite Background**

Digital Broadcast Satellite (DBS) technology provides an asymmetrical uplink/downlink bandwidth broadcast network, similar to the CATV network described above. The technology relies on relatively high power, widely-spaced  
30 satellites which can achieve digital transmission rates of either 30 Mbps (in a reduced error-correcting code mode of 30 Mbps for data and 10 Mbps for protection coding) or 23 Mbps (in a more robust coding configuration with 23 Mbps for data

and 17 Mbps for protection coding). Either configuration is achieved with a digitally coded 23 MHz broadband satellite channel.

Use of the channel for television signals is obtained with digital compression technologies, either the Motion Picture Experts Group (MPEG) standard (MPEG-II) or proprietary standards developed by industry. There is heavily reliance on complex real-time decompression technologies embedded in the receiving units, which can be viewed as the DBS analogue of the set-top-box in the CATV domain. The receiving units are characterized by small receiving dishes (18"-36"), connected to the set-top-box analogue via coaxial cabling or other means. Decompression algorithms operant in the box convert the raw digital signals to a format suitable for display on an NTSC television. The algorithms are of considerable complexity and are implemented via Application-Specific Integrated Circuit (ASIC) or Digital Signal Processor type technologies in combination with other specialized processing. The total cost of configurations can be brought to under \$1000 dollars. The cost of the electronic components will decline consistent with other electronic communication and computation technologies, although limits defined by power and packaging will remain and limit the cost reductions possible.

The memory requirements of any configurations are expected to be small, as the television applications are the main application envisioned for the technology. Many of the proposed systems (e.g., PrimeStar, DIRECTV/USS, EchoStar and AlphaStar) are television-oriented, although there is some direct access to the digital streams provided via high-speed access ports in some implementations of the set-top-box analogues. While this seems intended for future High-Definition Television (HDTV) applications, its presence illustrates a considerable opportunity to use this high-speed digital access in future configurations. The set-top-box analogues typically provide an RJ11 jack used for communication with the Public-Switched Telephone Network (PSTN) which is used for transmission of billing information or so-called "impulse pay-per-view" response information.

This data distribution system is comprised of clients, which are end-user terminals with extremely limited storage capacity (dictated by the factors we have discussed previously in this disclosure), and servers, which contain multimedia

information (such as video-on-demand) and in addition descriptive information (directory information, schedules, indices) for the multimedia information. Clients are assumed to be under customer physical control, while servers in the DBS case are likely to be uplinked to the satellite transponder. This results in a distributed  
5 architecture for directory information which is robust in the face of change, is low-cost, and utilizes the DBS infrastructure itself to preserve these properties. In practice, the client of the architecture is the "set-top box" like entity for DBS that we have previously discussed.

### **Special Considerations for DBS Versus CATV**

10 There are a number of limitations of the DBS system versus CATV. In particular, these are:

1. Error rates - The characteristics of the radio channel are such that heavyweight error protection coding must be in place, consuming up to 50% of channel capacity. This is a major difference, but does not affect our architecture.
- 15 2. Low exploitation of Hierarchy (extremely flat network) A major problem with DBS or satellite is general is that it is broadcast and wide area. Broadcast is not so much of a problem as the wide area. Broadcast is a specialization of multicast, and de facto, means that all interested receivers will find the information available. The wide area advantage is that the satellite material can be widely  
20 distributed, but does not give us the opportunity to insert memory nodes at points internal to the multicast distribution tree, as in fact there are NO such points.
3. Much lower back channel bandwidth - CATV offers the possibility of reasonably high performance channels which clients can use to select cells for inclusion in their memory regions. Required transmitter power precludes using a  
25 satellite transceiver at client sites; there is a receiver and another means of reply (e.g., the PSTN).
4. Interoperability.- There is possibly a limited ability to interoperate with the ATM cell-like format we have utilized in this disclosure. An overlay is possible on the underlying digital packet format using appropriate fragmentation and  
30 reassembly algorithms at the boundary between formats.

### **Memory Reduction Technique In Wireless ATM Technology**

Asynchronous Transfer Mode (ATM) is a link multiplexing technique which has found wide use in core telecommunications networks ("backbones") due to its ability to allocate bandwidth resources at a fine granularity, e.g., 48-octet "cells" of data. Considerable background material on ATM exists in the technical literature; for example A. Fraser of Bell Laboratories discusses the history of the Asynchronous Time-Division Multiplexing Technique in his 1994 ACM SIGCOMM Award lecture, which has been reprinted in ACM Computer Communication Review. There are a number of publications which discuss terminal adapters for these networks in the context of high-performance workstations. These designs require considerable memory to operate successfully, mainly for buffering of arriving data in order to prevent dropped data, and its consequence of poor system performance. Considerably less attention has been paid to ATM support for lower-cost devices, such as those which may used by consumers for everyday tasks in the future, e.g., wireless personal digital assistants. These devices will be characterized by low cost, light weight, and will place a premium on long battery life. Wireless ATM is operational in laboratory settings and can bear video traffic. Thus, it provides an environment to which the memory-reduction schemes we developed for CATV and DBS also apply.

### **Memory Reduction For Mobile Viewing Systems**

Reducing the memory requirements of future wireless terminals is very important today for technology and cost reasons, and is likely to remain so in the foreseeable future. Memory is the major cost in many of today's information handling devices, tethered or wireless. Memory is a major consumer of power in mobile wireless systems, and that a consequence is that reduced memory results in longer battery life in many cases. The advantages of memory-use reduction techniques in wireless systems are summarized as follows:

1. The systems are more compact, as less memory and fewer batteries are needed.
2. For the same reasons, they are lighter.
3. Longer battery life for diskless mobile terminals can be achieved.
4. They cost less.

### **Implementation Of A Client/Server Architecture On Wireless ATM**

The data distribution system is comprised of clients, which are end-user terminals with extremely limited storage capacity (dictated by the factors we have discussed previously in this disclosure), and servers, which contain multimedia information (such as video-on-demand) and in addition descriptive information (directory information, schedules, indices) for the multimedia information. Clients are assumed to be under customer physical control. This results in a distributed architecture for directory information which is robust in the face of change, is low-cost, and utilizes the wireless ATM infrastructure. In practice, the client of the architecture is a low-cost mobile "set-top box" analogue, whose advantages we have summarized briefly in Section 2.

### **Implementation of the Client/Server Architecture**

The data distribution system is comprised of clients, which are end-user terminals with extremely limited storage capacity (dictated by the factors we have discussed previously in this disclosure), and servers, which contain multimedia information (such as video-on-demand) and in addition descriptive information (directory information, schedules, indices) for the multimedia information. Clients are assumed to be under customer physical control, while servers in the DBS case are likely to be uplinked to the satellite transponder. This results in a distributed architecture for directory information which is robust in the face of change, is low-cost, and utilizes the DBS infrastructure itself to preserve these properties. In practice, the client of the architecture is the "set-top box" like entity for DBS that we have previously discussed.

There are several alternatives for our Client/Server architecture to be realized. Among these are:

1. sharing of a dish among several client sites, allowing a limited form of hierarchy and memory savings. A very simple example of this might be in an apartment building or townhouse configuration.

2. storing of considerable data in memory at the uplink point in the system.

This is fact where the server portion of the architecture should be located for our architecture to work. The idea here is that the server site contains all of the information that the small-memory set top boxes might need. Without hierarchy,

the selectivity of the similarity algorithms will not offer any statistical gain. The method to attack this problem is as follows:

1. Allocate a fraction of a DBS channel. This channel will be used as a "memory" by the client/server algorithm. For example, consider using 1 Mbps of bandwidth. This would give about 2300 cells per second. The use of the similarity algorithm would be to select which cells to send in the next second. If the response time must be better than 1 second, the scheduling interval can be a smaller value, such as 1/30 second (although round-trip delays could be problematic).

2. Schedule the bandwidth utilization based on the need for memory cells at the client. This schedule is VERY similar to that used in the CATV-oriented realization.

#### **Server/Ground Station Algorithm:**

The overall algorithm used by the server is as follows:

1. If new directory information is loaded into the server, the server stores the information in a schedule data structure. If this information represents an update of previous information, the recipients of the previous information (the "copy set") are informed that they should delete that information. This ensures that while not necessarily complete, the information at each client will be correct.

2. If the server receives a request from a client for directory information, it obtains the most current information from its own storage and sends this to the client.

3. The server periodically sends a time-of-day and date indicator to the client system.

#### **Detailed Description of Server Actions**

Each head end 100 in the broadcast or multicast network operates in a manner as noted above to distribute a limited subset of directory information to the subscriber terminal adapters (clients) 112 to thereby enable the subscribers to review only the segments of the overall directory information which are pertinent to that subscriber. To accomplish this task in an expeditious fashion, the format of the data, its content and management are regulated to ensure data transmission and storage efficiency. Therefore, the following list of characteristics represents a

specific implementation of this data distribution architecture for the CATV application:

1. The head end 100 is equipped with a clock which maintains the current time of day.
- 5        2. The head end 100 has a current "schedule" for directory information delivery. Directory information is indexed by <Channel Identifier, Slot Start, Slot End> 3-tuples. These indices are used to look up information about the program material presented in the identified slot.
- 10       3. A Slot is typically a predetermined segment of time, such as a ½ hour time period, while a Channel Identifier is data representative of the broadcast channel number on the broadcast network. This architecture also applies to video-on-demand starting at arbitrary times, with arbitrary durations, or minute by minute "microscheduling" of program material, since the slot start and end times are noted as well as the channel on which the subscriber terminal adapter can access the  
15       desired program material.
- 20       4. The program material which is broadcast in the identified time slot is described by a second 3-tuple: <Name, Textual Description, Other>. The Name field is the name used for the program material and is typically a short piece of text. The Textual Description field is text describing the program material, and might include summaries, movie reviews, script excerpts, actors names, awards won by the film, etc. The Other field is intended for extensibility at this point in time, and might include still photos, indications of violence or other adult content, or information about other program material in a series.
- 25       5. Program Material is stored in memory in the head end 100 in fixed-size units. The extent of a fixed-size unit is selected to minimize wasted space in both head end 100 and terminal adapter 112. In a memory architecture with a fixed size blocks, the storage of large variable-size objects (the descriptions) typically results in ½ of the block size being wasted. Such considerations have been extensively discussed in the literature on demand-paged virtual memory. For this disclosure,  
30       the storage units are termed "cells."
6. While not restricting the cell size to 48 octets (bytes), this size may be useful for interoperability with future computer networks based on the Asynchronous

WO 98/57696  
Transfer Mode (ATM) technology, which provides cells (packets) with a payload size of 48 octets.

7. A cell size of 48 bytes should capture most program material Titles in a single cell. The techniques described next can be used if the Title information is  
5 longer than a single cell.

8. The Textual Description and Other information are variable-sized and in all likelihood larger than a single cell. They are easily stored using multiple cells, either by allocating a sequence of cells large enough to contain the complete information from the memory of head end 100, or alternatively, allocating a linked  
10 list of cells where a cell contains information leading to the next cell. An example of such a linked list in the "C" computer programming language providing the appropriate structure would be:

```
struct linked_list_of_cell_bodies { char *c_body;  
/* pointer to 48 byte cell */  
15 struct linked_list_of_cell_bodies *c_next;    };
```

Therefore, the text "and Peter Sellers plays both the President and Doctor Strangelove in this classic black comedy about the Cold War" might be stored in three pieces: "and Peter Sellers plays both the President and D," in Cell Body #1, "Doctor Strangelove in this classic black comedy a," in Cell Body #2, and "bout the  
20 Cold War" in Cell Body #3.

The linked list data structure, like a more complex virtual memory scheme, offers advantages in multiplexing storage in comparison to a scheme where contiguous areas of memory are required.

9. The head end 100 periodically transmits directory information to clients  
25 along the downlink channel. A key innovation of this data distribution system is the method for scheduling the transmission of directory information. It combines the similarity-measurement technology with specific characteristics of the high-bandwidth multicast downlink/low-bandwidth unicast uplink networks. Information is transmitted in units of cells, and for specificity consider these cells to be 48 octet  
30 data objects. At any data rate, a set of number of cells per unit time, e.g., for a MHz channel, one bit per baud, there are 6 megabits or 750,000 octets per second transmitted. These octets can be grouped as 48 octet cells and 15,625 such cells



can be transmitted per second. While any interval can be used, it is convenient to use an interval of 1 second, as this is both a convenient term for engineering analyses and happens to be an acceptable delay for most human subscribers to consider behavior-"instantaneous." With 15,625 cells per second, each second in the schedule could contain, for example, titles for all current programs on a 500 channel system and leave 15,000 cells for other information in the same second. It should be clear that higher data rates provide more cells per second and lower data rates provide fewer.

The schedule for the cells, Cell[1]...Cell[15625] is developed using the novel schemes of our previous application combined with some new features to reduce memory usage and support rapid interaction. The scheduling problem can be looked at as a job-shop scheduling problem. Subscriber requests for information (described below when discussing the Client Algorithm) can be viewed as requests for transmission of cells. These cells might contain title information, Textual Descriptions of program material, or Other information about the program material. Each cell is fully identified in the system as a whole by a <Channel Identifier, Time, Cell #> 3-tuple, with the Cell # indicating its position in the <Name, Textual Description, Other> 3-tuple. These indices can be made more or less structured to ease implementation.

10. The schedule is developed according to priorities from highest to lowest. An example priority scheme is:

- a.) Subscriber requests for Names, oldest request first, one cell interval at a time is allocated until all the requested Names are transmitted.
- b.) Subscriber requests for Textual Descriptions, oldest request first. For equally old requests, one cell interval at a time is allocated until all the requested text is transmitted.
- c.) Subscriber requests for Other information, oldest request first. For equally old requests, one cell interval at a time is allocated until all the requested data is transmitted.
- d.) Textual Descriptions associated with subscriber Name requests, oldest request first. For equally old requests, one cell interval at a time is allocated until all the requested text is transmitted.

e.) Other information associated with subscriber Name requests, oldest request first. For equally old requests, one cell interval at a time is allocated until all the requested text is transmitted.

5 f.) Names for program material in future time intervals determined similar to subscriber interests based on analysis of textual descriptions or other subscriber behavior available to the scheduler, such as typical usage times for the viewing device, typical channel selections, etc.

g.) Textual descriptions,....

h.) Other,...

10 11. The intent of the priority scheme is to preserve good interactive behavior, fairness to the subscriber population (sharing of the channel they all pay for) and good utilization of the downstream channel. Giving lower priorities to the anticipated viewing habits allows the information to be preloaded to the subscriber terminal adapters when the channel is otherwise unneeded by specific subscriber  
15 requests. When subscribers have made specific requests, the requests take priority in order to provide best interactive response time. Names take priority as Names are often used to channel surf, and bandwidth would be wasted if the descriptions were sent for Names that are discarded. Other priority schemes can be used, such as pricing-based or closeness to the present time and the present  
20 example is merely meant to illustrate and teach how a schedule for cell transmission at a Server can be developed under a set of policy constraints.

#### **Client Algorithm**

In an analogous manner, the following list of characteristics represents a specific implementation of the algorithm used by the Client (subscriber terminal  
25 adapter 112) in this data distribution architecture for the CATV application:

1. Client memory is organized in units of cells.
2. Clients have limited memory.
3. Clients prioritize the storage of information.
4. One priority scheme might be:

30 a.) Name of material presently being viewed  
b.) Textual Description of material being viewed, or even only the first few cells of this if memory is extremely limited;

c.) Other information about the material currently being viewed, or even only the first few cells of this if memory is extremely limited;

d.) Name information of material presently available on a selected set of other channels;

5 e.) Textual Descriptions and Other information on material currently available on a selected set of other channels;

f.) Name information of material available in the future on a selected set of other channels, prioritized by nearness in time;

10 g.) Textual Descriptions and Other information on material available in the future on a selected set of other channels, prioritized by nearness in time.

Presuming sufficient storage is available, affinity matrices can be developed to select channel and program material. The usage of the Client Storage is optimized when the subscriber terminal adapter is able to provide the subscriber.  
15 instantaneous or near-instantaneous selection of material with a minimum of data storage. Of course, as with most information systems, more data storage improves performance, but even the most minimal data storage architecture (1 cell) combined with the Client/Server protocol, can provide a correctly functioning data distribution system. A single 64 Kbit RAM chip provides capacity for slightly more than 160  
20 cells, so this extreme is unlikely. When the Server correctly anticipates Client needs, the downlink schedule is optimized.

5. Information for expired time intervals is immediately discarded.

6. Information from the server is requested as <Channel Identifier, Time, Cell #> 3-tuples.

25 7. Information requests are spurred by subscriber interaction, or expected needs based on history data maintained at the server.

8. Information requests can include requests to send groups of cells, such as those associated with a complete Textual Description or Other information. This allows the multiplexing of the downstream channel  
30 between subscribers as described in the Server algorithm section.

9. When a cell arrives, for example after a request has been issued, it is examined for priority versus the present set of cells. If there are available

unused cells in the memory of the subscriber terminal adapter, the received cell is stored. If no cells are free in the memory of the subscriber terminal adapter, the priority of the cell is compared to cells presently stored in the subscriber terminal adapter. If the priority of the newly received cell is higher than the priority of at least one previously stored cell, the lowest priority resident cell is discarded and replaced by the higher-priority new arrival cell. Note that this process applies to both requested and unrequested cells and allows for anticipated usage when memory is available to do so, as well as cell-by-cell updates for extremely low-capacity subscriber terminal adapters.

## 10 **Prefetch**

Bandwidth is greater at the root end of a hierarchical node network system, in anticipation of a request, it is prudent to use the similarity measure to predictively cache files into local servers and further to narrowcast selections into given distribution cells and (hierarchical) sub-cells through sub-servers based on what selections are most likely to be requested in each cell so as to significantly increase the utilization of bandwidth via this hierarchical narrowcast configuration. The importance of this savings appears in proportion to the degree of granularity (smallness of the cell) in the narrowcast architecture. This technique can also be used to make decisions for scheduling what data should be placed on dedicated channels. This may be more network efficient if a file were popular enough to be continuously in the queue because upon submission of a request a file may be partially downloaded regardless of where in the length of the file the download began. The (initially) missed portion of the file can then be immediately picked up as the file narrowcast starts over, thus completing the download in the same time period as if a special request for the file were made. Mobile users whose geographic locations are known can have files precached (e.g. at night) into the servers which are presently physically closest to them at any given time

The present data distribution system employs the idea of prefetching, which has also been referred to as pre-caching, cache-preloading, or anticipation in the technical literature. The basic idea is that if good predictions of future data requirements are available, and there is excess data-fetching capability available, the data should be fetched aggressively in anticipation of future needs. If

successful, this technique has two major benefits applicable to present and future networks. First, it can reduce (i.e., improve) response-time, a major performance advantage in interactive systems. Second, it can reduce congestion and other problems associated with network overload. In Figure 8, we illustrate a possible  
5 time series of bandwidths required by an application or set of applications. To understand how the responsiveness of the system is improved, it can be seen that the unused bandwidth can be used to transmit information likely to be used in the future. For example, if a list is being traversed 1,2,3,4,... it is likely that if object N has been requested, that object N+1 will be the next request. If N+1 is prefetched  
10 from the remote system, it will be available when the request is made with additional delays. All of the "UNUSED BANDWIDTH" illustrated in Figure 8 can potentially be used to pre-fetch. To understand why congestion can be reduced, we first note that congestion is due to inadequate resources. The simplest example is in sufficient road capacity for the offered load, leading to "traffic jams." Packet-  
15 switched networks such as the IP Internet suffer from similar problems. In Figure 8, it appears that capacity is being reached or exceeded near TIME=43 or so. If demands are made during this period, traffic is likely to be delayed or even lost. If we prefetch successfully during more lightly-loaded periods (such as TIME=0..42), we reduce the probability of data being requested in the future, essentially trading  
20 the guarantee of a fully loaded network today for the promise of no congestion in the future. By fetching data in anticipation of future needs, we reduce (at least probabilistically) those future needs.

Prefetching has been used in the computer operating systems field for several decades, and a variety of algorithms have been explored. A. J. Smith of  
25 Berkeley has reported that the only case where successful predictions about future requests for memory objects can be made is when accesses are sequential. More recent work for higher-level content such as World-Wide Web (WWW) hypertext has shown that user-authored links to other hypermedia documents can be followed with some success.

30 The prefetching technology of this disclosure is based on unused slots being filled with pre-sent information based on our understanding of user interest using the similarity measures developed in the large original application, and used for our

prioritization (see p. 18 of the current invention disclosure), a concept that they do not deal with, as they follow http: links based on observations about the high probability that these links will be followed by users.

The present invention may be usefully applied within the context of set-top  
5 box like devices such as personal digital assistants or network computers, or personal computers used as a form of set-top box, as a method of reducing response time as observed by users. This method used "links" to other documents embedded in an HTTP-format file as hints that those links should be followed in prefetching data; that is, the linked documents should be fetched in anticipation of  
10 the user's desire to follow the links to those documents.

The present invention provides two enhancements to this scheme. First, it provides a technological means by which the prefetched data can be intermixed with on-demand data to provide overall improvements in response time to a large population of HTTP/WWW users, with reduced memory requirements. Second, the  
15 present invention, which views the downlink as a fixed capacity resource, provides a general scheduling method embodying techniques such as user preferences to prefetch when slots or bandwidth are underutilized, to preemptively reduce future demand for bandwidth. The general technique of using similarity-informed prefetching is described at length in a copending patent application entitled  
20 "Pseudonymous Server For System For Customized Electronic Identification Of Desirable Objects", U.S. Patent Application Serial No. 08/550,886, filed October 31, 1995.

### **Artificial Delays**

This note discusses the notion of "artificial delays" in the queuing of requests  
25 for the satellite or cable system to which our set top boxes are attached. The idea is that by careful management of the queues, we can effect significant bandwidth savings for the system as a whole. If you will remember, the Server scheduling algorithm (I've attached the text for it from the DBS scheme I sent during the Summer at the end of this e-mail) goes like this:

30 The client set-top box (of which there should be many) sends REQUESTs for information in cell-sized units to the server system. The server system applies a priority algorithm (see especially Step 10, below) to CHOOSE the next cell to

send. By design of the relative priorities, we can get good responsiveness and reduced bandwidth needs, in spite of the low memory needs (and low cost) of the set top boxes.

Imagine the scenario where there are MANY set tops connected to the server. This situation might be as illustrated in Figure 9, where the Cs are Clients and the S is a Server. Now clearly, there is a multiplicity of Clients, and by virtue of this multiplicity, we may be able to achieve a savings through appropriate delays. I again believe the similarity measure is the key to success here, and to novelty. Consider the cell requests for Clients C1, C4 and C5, shown below using letters to indicate particular cells as discussed in our disclosure text.

C1: E-T-A-O-I-N-S-H-R-D-L-U.....

C4: N-A-T-I-O-N-A-L-V-E-L.....

C5: E-A-T-O-N-L-Y-S-U-D.....

15

We mark these cell request with times associated with their transmission intervals:

T:

I:0 0 0 0 0 0 0 0 1 1 1 1

20

M:1-2-3-4-5-6-7-8-9-0-1-2-3.....

E:

Now, for convenience, assume that all of the cell requests show above have the same priority. Then the server might actually send the following sequence of cells over the channel:

25

S:E-N-E-T-A-A-A-T-T-O-I-O.....

Thus, we are servicing the cell requests C1-C4-C5-C1-C4-C5..... (in fact, the Server may notice the overlaps between requests by C1 and C5 in the first interval, C4 and C5 in the second interval, C4 and C5 in the third interval, and C1 and C5 in the fourth interval, giving:

30

S:E-N-T-A-A-T-O-I-I-O-N-N...

)

Imagine that the clients are always listening. Then, we can delay cell requests in  
 5 the HOPE that the REPLYs can be MERGED, satisfying multiple set-top-box  
 clients with the same REPLY. To make this concrete, consider delaying service by  
 one period. So the output of the server then looks like:

S:#-E-N-T-A-O-I-N-L-S-A-Y....

10

What is going on here is very subtle. By delaying some clients service requests, we  
 are INCREASING THE PROBABILITY that another such request will come in,  
 which can we can fold into service of the equivalent delayed request. The cost is  
 potentially in delay, but with enough overlap, the cell times are short enough for 48  
 15 bytes on a DBS channel that we can probably delay significantly.

Considering the problem theoretically for a moment, we can compute the  
 gain for the an acceptable delay D as being the number of redundant transmissions  
 which are eliminated due to a delay D. So, for delays of from 1 to 10 cell times the  
 total number of DBS cells without redundancy checks is 30; the number required  
 20 when this small optimization is applied is as shown below:

Delay DBS CellsBandwidth Savings

	02430/24=25%
25	11830/18=66%
	21730/17=76%
	31730/17=76%
	41530/15=100%
	51530/15=100%
30	61530/15=100%
	71530/15=100%
	81530/15=100%



$$91530/15=100\%$$

$$101430/14=114\%$$

We compute the bandwidth gain against the dumb use of 30 cell times; the  
5 bandwidth gain comes from the fact that the synchronous satellite channel gives  
us a fixed bandwidth, giving a fixed number of cells per unit time, and we have just  
saved 16 cell times by use of the delay scheme. For this example, at this point no  
more gain is possible, since all the duplication has been eliminated. In some sense,  
this behaves like a compression scheme. The similarity algorithm increases the  
10 probability that these overlaps will occur - the ideal situation is we are waiting long  
enough so that the scheduled broadcast cell satisfies almost all requests for that  
cell within a significant time intervals (say several milliseconds).

#### **Basic Filtering of Programs**

The data distribution system of the present invention automatically  
15 constructs both a target profile for each target object in the electronic media, as well  
as a "subscriber target profile interest summary" for each subscriber, which  
subscriber target profile interest summary describes the subscriber's interest level  
in various types of target objects. The system then evaluates the target profiles  
against the subscribers' target profile interest summaries to generate a subscriber-  
20 customized rank ordered listing of target objects most likely to be of interest to each  
subscriber so that the subscriber can select from among these potentially relevant  
target objects, which were automatically selected by this system from the plethora  
of target objects available on the electronic media.

Because people have multiple interests, a target profile interest summary for  
25 a single subscriber must represent multiple areas of interest, for example, by  
consisting of a set of individual search profiles, each of which identifies one of the  
subscriber's areas of interest. Each subscriber is presented with those target  
objects whose profiles most closely match the subscriber's interests as described  
by the subscriber's target profile interest summary. Subscribers' target profile  
30 interest summaries are automatically updated on a continuing basis to reflect each  
subscriber's changing interests. In addition, target objects can be grouped into  
clusters based on their similarity to each other, for example, based on similarity of

their topics in the case where the target objects are published programs, and menus automatically generated for each cluster of target objects to allow subscribers to navigate throughout the clusters and manually locate target objects of interest. For reasons of confidentiality and privacy, a particular subscriber may not wish to make public all of the interests recorded in the subscriber's target profile interest summary, particularly when these interests are determined by the subscriber's purchasing patterns. The subscriber may desire that all or part of the target profile interest summary be kept confidential. It is therefore necessary that data in a subscriber's target profile interest summary be protected from unwanted disclosure except with the subscriber's agreement, which may entail processing at least part of the agreement matrix in the subscriber terminal adapter. At the same time, the subscriber's target profile interest summaries must be accessible to the relevant servers that perform the matching of target objects to the subscribers, if the benefit of this matching is desired by both providers and consumers of the target objects.

In the preferred embodiment of the invention, the data distribution system uses a fundamental methodology for accurately and efficiently matching subscribers and target objects by automatically calculating, using and updating profile information that describes both the subscribers' interests and the target objects' characteristics. The target objects are typically programs, and their properties are stored, and/or represented and /or denoted on the electronic media as (digital) data. The information delivery process in the preferred embodiment is based on determining the similarity between a profile for the target object and the profiles of target objects for which the subscriber (or a similar subscriber) has provided positive feedback in the past. The individual data that describe a target object and constitute the target object's profile are herein termed "attributes" of the target object. Attributes may include, but are not limited to, the following: (1) long pieces of text (a movie review, or an advertisement), (2) short pieces of text (name of a movie's director, name of town from which an advertisement was placed), (3) numeric measurements (rating given to a movie), (4) associations with other types of objects (list of actors in a movie). Any of these attributes, but especially the numeric ones, may correlate with the quality of the target object, such as measures

of its popularity (how often it is accessed) or of subscriber satisfaction (number of complaints received).

Each movie has a different set of values for these attributes. This example conveniently illustrates three kinds of attributes. It is evident that they can be used to help the subscriber identify target objects (movies) of interest. For example, the subscriber might previously have rented many Parental Guidance (PG) films, and many films made in the 1970's. This generalization is useful: new films with values for one or both attributes that are numerically similar to these (such as MPAA rating of 1, release date of 1975) are judged similar to the films the subscriber already likes, and therefore of probable interest. Textual attributes are important for helping the subscriber locate desired films. For example, perhaps the subscriber has shown a past interest in films whose review text contains words like "chase," "explosion," "explosions," "hero," "gripping," and "superb." This generalization is again useful in identifying new films of interest. The TF/IDF method can be used to profile the movie descriptions, or even the closed captioning contained therein, as described below. Associative attributes record associations between the target objects in this domain, namely movies, and ancillary target objects of an entirely different sort, namely humans.

#### **Decomposing Complex Attributes**

Although textual and associative attributes are large and complex pieces of data, for information retrieval purposes they can be decomposed into smaller, simpler numeric attributes. This means that any set of attributes can be replaced by a (usually larger) set of numeric attributes, and hence that any profile can be represented as a vector of numbers denoting the values of these numeric attributes. In particular, a textual attribute, such as the full text of a movie review, can be replaced by a collection of numeric attributes that represent scores to denote the presence and significance of the words in that text. The score of a word in a text may be defined in numerous ways. The simplest definition is that the score is the rate of the word in the text, which is computed by computing the number of times the word occurs in the text, and dividing this number by the total number of words in the text. This sort of score is often called the "term frequency" (TF) of the word. The definition of term frequency may optionally be modified to

weight different portions of the text unequally: for example, any occurrence of a word in the text's title might be counted as a 3-fold or more generally k-fold occurrence (as if the title had been repeated k times within the text), in order to reflect a heuristic assumption that the words in the title are particularly important indicators of the text's content or topic.

Just as a textual attribute may be decomposed into a number of component terms (letter or word n-grams), an associative attribute may be decomposed into a number of component associations. For instance, a typical associative attribute used in profiling a movie would be a list of subscribers who have rented that movie. This list can be replaced by a collection of numeric attributes, which give the "association scores" between the movie and each of the subscribers known to the system. For example, the 165th such numeric attribute would be the association score between the movie and subscriber #165, where the association score is defined to be 1 if subscriber #165 has previously rented the movie, and 0 otherwise. In a subtler refinement, this association score could be defined to be the degree of interest, possibly zero, that subscriber #165 exhibited in the movie, as determined by relevance feedback.

### Similarity Measures

What does it mean for two target objects to be similar? More precisely, how should one measure the degree of similarity? Many approaches are possible and any reasonable metric that can be computed over the set of target object profiles can be used, where target objects are considered to be similar if the distance between their profiles is small according to this metric. Thus, the following preferred embodiment of a target object similarity measurement system has many variations.

First, define the distance between two values of a given attribute according to whether the attribute is a numeric, associative, or textual attribute. If the attribute is numeric, then the distance between two values of the attribute is the absolute value of the difference between the two values. When necessary for clustering or other purposes, a metric that allows comparison of any two target objects (whether of the same or different sorts) can be defined as follows. If  $a$  is an attribute, then let  $\text{Max}(a)$  be an upper bound on the distance between two values of attribute  $a$ ;

notice that if attribute  $a$  is an associative or textual attribute, this distance is an angle determined by  $\arccos$ , so that  $\text{Max}(a)$  may be chosen to be 180 degrees, while if attribute  $a$  is a numeric attribute, a sufficiently large number must be selected by the system designers. The distance between two values of attribute  $a$  is given as before in the case where both values are defined; the distance between two undefined values is taken to be zero; finally, the distance between a defined value and an undefined value is always taken to be  $\text{Max}(a)/2$ . This allows us to determine how close together two target objects are with respect to an attribute  $a$ , even if attribute  $a$  does not have a defined value for both target objects.

10 The distance  $d(*,*)$  between two target objects with respect to their entire multi-attribute profiles is then given in terms of these individual attribute distances exactly as before. It is assumed that one attribute in such a system specifies the sort of target object ("movie", "novel", etc.), and that this attribute may be highly weighted if target objects of different sorts are considered to be very different despite any

15 attributes they may have in common.

#### **Filtering: Relevance Feedback**

A filtering system is a device that can search through many target objects and estimate a given subscriber's interest in each target object, so as to identify those that are of greatest interest to the subscriber. The filtering system uses

20 relevance feed back to refine its knowledge of the subscriber's interests: whenever the filtering system identifies a target object as potentially interesting to a subscriber, the subscriber (if an on-line subscriber) provides feedback as to whether or not that target object really is of interest. Such feedback is stored long-term in summarized form, as part of a database of subscriber feedback information,

25 and may be provided either actively or passively. In active feedback, the subscriber explicitly indicates his or her interest, for instance, on a scale of -2 (active distaste) through 0 (no special interest) to 10 (great interest). In passive feedback, the system infers the subscriber's interest from the subscriber's behavior. For example, if target objects are textual documents, the system might monitor which documents

30 the subscriber chooses to read, or not to read, and how much time the subscriber spends reading them.

**Filtering: Determining Topical Interest Through Similarity**

Relevance feedback only determines the subscriber's interest in certain target objects: namely, the target objects that the subscriber has actually had the opportunity to evaluate (whether actively or passively). For target objects that the subscriber has not yet seen, the filtering system must estimate the subscriber's interest. This estimation task is the heart of the filtering problem, and the reason that the similarity measurement is important. The subscriber can provide active and/or passive feedback to the system relating to presented programs. However, the system does not have feedback information from the subscriber for programs that have never been presented to the subscriber, such as new programs that have just been added to the broadcast schedule, or old programs that the system chose not to present to the subscriber.

As shown in flow diagram form in Figure 6, the evaluation of the likelihood of interest in a particular target object for a specific subscriber can automatically be computed. The interest that a given target object  $X$  holds for a subscriber  $U$  is assumed to be a sum of two quantities:  $q(U, X)$ , the intrinsic "quality" of  $X$ , plus  $f(U, X)$ , the "topical interest" that subscribers like  $U$  have in target objects like  $X$ . For any target object  $X$ , the intrinsic quality measure  $q(U, X)$  is easily estimated at steps 601-603 directly from numeric attributes of the target object  $X$ . The computation process begins at step 601, where certain designated numeric attributes of target object  $X$  are specifically selected, which attributes by their very nature should be positively or negatively correlated with subscribers' interest. Such attributes, termed "quality attributes," have the normative property that the higher (or in some cases lower) their value, the more interesting a subscriber is expected to find them. Quality attributes of target object  $X$  may include, but are not limited to, target object  $X$ 's popularity among subscribers in general, the rating a particular reviewer has given target object  $X$ , the age (time since authorship -- also known as outdatedness) of target object  $X$ , the number of vulgar words used in target object  $X$ , the price of target object  $X$ , and the amount of money that the company selling target object  $X$  has donated to the subscriber's favorite charity. At step 602, each of the selected attributes is multiplied by a positive or negative weight indicative of the strength of subscriber  $U$ 's preference for those target objects that have high

values for this attribute, which weight must be retrieved from a data file storing quality attribute weights for the selected subscriber. At step 603, a weighted sum of the identified weighted selected attributes is computed to determine the intrinsic quality measure  $q(U, X)$ . At step 604, the summarized weighted relevance

5 feedback data is retrieved, wherein some relevance feedback points are weighted more heavily than others and the stored relevance data can be summarized to some degree, for example by the use of search profile sets. The more difficult part of determining subscriber U's interest in target object X is to find or compute at step 605 the value of  $f(U, X)$ , which denotes the topical interest that subscribers like U

10 generally have in target objects like X. The method of determining a subscriber's interest relies on the following heuristic: when X and Y are similar target objects (have similar attributes), and U and V are similar subscribers (have similar attributes), then topical interest  $f(U, X)$  is predicted to have a similar value to the value of topical interest  $f(V, Y)$ . This heuristic leads to an effective method because

15 estimated values of the topical interest function  $f(*, *)$  are actually known for certain arguments to that function: specifically, if subscriber V has provided a relevance-feedback rating of  $r(V, Y)$  for target object Y, then insofar as that rating represents subscriber V's true interest in target object Y, we have  $r(V, Y) = q(V, Y) + f(V, Y)$  and can estimate  $f(V, Y)$  as  $r(V, Y) - q(V, Y)$ . Thus, the problem of estimating

20 topical interest at all points becomes a problem of interpolating among these estimates of topical interest at selected points, such as the feedback estimate of  $f(V, Y)$  as  $r(V, Y) - q(V, Y)$ . This interpolation can be accomplished with any standard smoothing technique, using as input the known point estimates of the value of the topical interest function  $f(*, *)$ , and determining as output a function that

25 approximates the entire topical interest function  $f(*, *)$ .

Not all point estimates of the topical interest function  $f(*, *)$  should be given equal weight as inputs to the smoothing algorithm. Since passive relevance feedback is less reliable than active relevance feedback, point estimates made from passive relevance feedback should be weighted less heavily than point estimates

30 made from active relevance feedback, or even not used at all. In most domains, a subscriber's interests may change over time and, therefore, estimates of topical interest that derive from more recent feedback should also be weighted more

heavily. A subscriber's interests may vary according to mood, so estimates of topical interest that derive from the current session should be weighted more heavily for the duration of the current session, and past estimates of topical interest made at approximately the current time of day or on the current weekday should be weighted more heavily. An estimate of topical interest  $f(V, Y)$  should be weighted more heavily if subscriber  $V$  has had more experience with target object  $Y$ . Indeed, a useful strategy is for the system to track long-term feedback for such target objects.

To effectively apply the smoothing technique, it is necessary to have a definition of the similarity distance between  $(U, X)$  and  $(V, Y)$ , for any subscribers  $U$  and  $V$  and any target objects  $X$  and  $Y$ . We have already seen how to define the distance  $d(X, Y)$  between two target objects  $X$  and  $Y$ , given their attributes. We may regard a pair such as  $(U, X)$  as an extended object that bears all the attributes of target  $X$  and all the attributes of subscriber  $U$ ; then the distance between  $(U, X)$  and  $(V, Y)$  may be computed in exactly the same way. This approach requires subscriber  $U$ , subscriber  $V$ , and all other subscribers to have some attributes of their own stored in the system: for example, age (numeric), social security number (textual), and list of documents previously retrieved (associative). It is these attributes that determine the notion of "similar subscribers." Thus it is desirable to generate profiles of subscribers (termed "subscriber profiles") as well as profiles of target objects (termed "target profiles"). Some attributes employed for profiling subscribers may be related to the attributes employed for profiling target objects: for example, using associative attributes, it is possible to characterize target objects such as  $X$  by the interest that various subscribers have shown in them, and simultaneously to characterize subscribers such as  $U$  by the interest that they have shown in various target objects. In addition, subscriber profiles may make use of any attributes that are useful in characterizing humans, such as those suggested in the example domain above where target objects are potential consumers. Notice that subscriber  $U$ 's interest can be estimated even if subscriber  $U$  is a new subscriber or an off-line subscriber who has never provided any feedback, because the relevance feedback of subscribers whose attributes are similar to  $U$ 's attributes is taken into account.



$$f(U, X) = \frac{\sum ((r(V, Y) - q(V, Y)) * g(\text{distance}(U, X) \wedge (V, Y)))}{\sum g(\text{distance}(U, V) \wedge (V, Y))}$$

For some uses of filtering systems, when estimating topical interest, it is appropriate to make an additional "presumption of no topical interest" (or "bias toward zero"). To understand the usefulness of such a presumption, suppose the system needs to determine whether target object X is topically interesting to the subscriber U, but that subscribers like subscriber U have never provided feedback on target objects even remotely like target object X. The presumption of no topical interest says that if this is so, it is because subscribers like subscriber U are simply not interested in such target objects and therefore do not seek them out and interact with them. On this presumption, the system should estimate topical interest  $f(U, X)$  to be low. Formally, this example has the characteristic that (U, X) is far away from all the points (V, Y) where feedback is available. In such a case, topical interest  $f(U, X)$  is presumed to be close to zero, even if the value of the topical interest function  $f(*, *)$  is high at all the faraway surrounding points at which its value is known. When a smoothing technique is used, such a presumption of no topical interest can be introduced, if appropriate, by manipulating the input to the smoothing technique. In addition to using observed values of the topical interest function  $f(*, *)$  as input, the trick is to also introduce fake observations of the form topical interest  $f(V, Y) = 0$  for a lattice of points (V, Y) distributed throughout the multidimensional space. These fake observations should be given relatively low weight as inputs to the smoothing algorithm. The more strongly they are weighted, the stronger the presumption of no interest.

The following provides another simple example of an estimation technique that has a presumption of no interest. Let  $g$  be a decreasing function from non-negative real numbers to non-negative real numbers, such as  $g(x) = e^{-x}$  or  $g(x) = \min(1, x^{-k})$  where  $k > 1$ . Estimate topical interest  $f(U, X)$  with the following  $g$ -weighted average:

Here the summations are over all pairs (V, Y) such that subscriber V has provided feedback  $r(V, Y)$  on target object Y, i.e., all pairs (V, Y) such that relevance feedback  $r(V, Y)$  is defined. Note that both with this technique and with

conventional smoothing techniques, the estimate of the topical interest  $f(U, X)$  is not necessarily equal to  $r(U, X) - q(U, X)$ , even when  $r(U, X)$  is defined.

**Filtering: Adjusting Weights and Residue Feedback**

The method described above requires the filtering system to measure  
5 distances between (subscriber, target object) pairs, such as the distance between  
(U, X) and (V, Y). Given the means described earlier for measuring the distance  
between two multi-attribute profiles, the method must therefore associate a weight  
with each attribute used in the profile of (subscriber, target object) pairs, that is, with  
each attribute used to profile either subscribers or target objects. These weights  
10 specify the relative importance of the attributes in establishing similarity or  
difference, and therefore, in determining how topical interest is generalized from  
one (subscriber, target object) pair to another. Additional weights determine which  
attributes of a target object contribute to the quality function  $q$ , and by how much.

It is possible and often desirable for a filtering system to store a different set  
15 of weights for each subscriber. For example, a subscriber who thinks of two-star  
films as having materially different topic and style from four-star films wants to  
assign a high weight to "number of stars" for purposes of the similarity distance  
measure  $d(*, *)$ ; this means that interest in a two-star film does not necessarily  
signal interest in an otherwise similar four-star film, or vice-versa. If the subscriber  
20 also agrees with the critics, and actually prefers four-star films, the subscriber also  
wants to assign "number of stars" a high positive weight in the determination of the  
quality function  $q$ . In the same way, a subscriber who dislikes vulgarity wants to  
assign the "vulgarity score" attribute a high negative weight in the determination of  
the quality function  $q$ , although the "vulgarity score" attribute does not necessarily  
25 have a high weight in determining the topical similarity of two films.

Attribute weights (of both sorts) may be set or adjusted by the system  
administrator or the individual subscriber, on either a temporary or a permanent  
basis. However, it is often desirable for the filtering system to learn attribute  
weights automatically, based on relevance feedback. The optimal attribute weights  
30 for a subscriber  $U$  are those that allow the most accurate prediction of subscriber  
 $U$ 's interests. That is, with the distance measure and quality function defined by  
these attribute weights, subscriber  $U$ 's interest in target object  $X$ ,  $q(U, X) + f(U, X)$ ,

can be accurately estimated by the techniques above. The effectiveness of a particular set of attribute weights for subscriber U can therefore be gauged by seeing how well it predicts subscriber U's known interests.

Formally, suppose that subscriber U has previously provided feedback on target objects  $X_1, X_2, X_3, \dots, X_n$ , and that the feedback ratings are  $r(U, X_1), r(U, X_2), r(U, X_3), \dots, r(U, X_n)$ . Values of feedback ratings  $r(*, *)$  for other subscribers and other target objects may also be known. The system may use the following procedure to gauge the effectiveness of the set of attribute weights it currently stores for subscriber U:

(i) For each  $1 \leq i \leq n$ , use the estimation techniques to estimate  $q(U, X_i) + f(U, X_i)$  from all known values of feedback ratings  $r$ . Call this estimate  $a_i$ .

(ii) Repeat step (i), but this time make the estimate for each  $1 \leq i \leq n$  without using the feedback ratings  $r(U, X_j)$  as input, for any  $j$  such that the distance  $d(X_i, X_j)$  is smaller than a fixed threshold. That is, estimate each  $q(U, X_i) + f(U, X_i)$  from other values of feedback rating  $r$  only; in particular, do not use  $r(U, X_i)$  itself. Call this estimate  $b_i$ . The difference  $a_i - b_i$  is herein termed the "residue feedback  $r_{res}(U, X_i)$  of subscriber U on target object  $X_i$ ."

(iii) Compute subscriber U's error measure,  $(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + \dots + (a_n - b_n)^2$ .

A gradient-descent or other numerical optimization method may be used to adjust subscriber U's attribute weights so that this error measure reaches a (local) minimum. This approach tends to work best if the smoothing technique used in estimation is such that the value of  $f(V, Y)$  is strongly affected by the point estimate  $r(V, Y) - q(V, Y)$  when the latter value is provided as input. Otherwise, the presence or absence of the single input feedback rating  $r(U, X_i)$ , in steps (i)-(ii) may not make  $a_i$  and  $b_i$  very different from each other. A slight variation of this learning technique adjusts a single global set of attribute weights for all subscribers, by adjusting the weights so as to minimize not a particular subscriber's error measure but rather the total error measure of all subscribers. These global weights are used as a default initial setting for a new subscriber who has not yet provided any feedback. Gradient descent can then be employed to adjust this subscriber's individual weights over time. Even when the attribute weights are chosen to minimize the error measure for subscriber U, the error measure is generally still positive, meaning that residue feedback from subscriber U has not been reduced to 0 on all

target objects. It is useful to note that high residue feedback from a subscriber U on a target object X indicates that subscriber U liked target object X unexpectedly well given its profile, that is, better than the smoothing model could predict from subscriber U's opinions on target objects with similar profiles. Similarly, low residue feedback indicates that subscriber U liked target object X less than was expected. By definition, this unexplained preference or dispreference cannot be the result of topical similarity, and therefore must be regarded as an indication of the intrinsic quality of target object X. It follows that a useful quality attribute for a target object X is the average amount of residue feedback  $r_{\text{res}}(V, X)$  from subscribers on that target object, averaged over all subscribers V who have provided relevance feedback on the target object. In a variation of this idea, residue feedback is never averaged indiscriminately over all subscribers to form a new attribute, but instead is smoothed to consider subscribers' similarity to each other. Recall that the quality measure  $q(U, X)$  depends on the subscriber U as well as the target object X, so that a given target object X may be perceived by different subscribers to have different quality. In this variation, as before,  $q(U, X)$  is calculated as a weighted sum of various quality attributes that are dependent only on X, but then an additional term is added, namely an estimate of  $r_{\text{res}}(U, X)$  found by applying a smoothing algorithm to known values of  $r_{\text{res}}(V, X)$ . Here V ranges over all subscribers who have provided relevance feedback on target object X, and the smoothing algorithm is sensitive to the distances  $d(U, V)$  from each such subscriber V to subscriber U.

#### **Compare Present Programs' Target Profile to a Subscriber's Target Profile Interest Summary**

The process by which a subscriber employs this apparatus to view programs is illustrated in flow diagram form in Figure 7. At step 701, the subscriber activates the subscriber terminal adapter 112. The subscriber has a subscriber target profile interest summary stored in distribution system 100. When the subscriber requests access to programs at step 702, the profile matching software resident on either distribution system 100 or in the subscriber terminal adapter 112 sequentially considers each search profile  $p_k$  from the subscriber's target profile interest summary to determine which programs are most likely of interest to the subscriber. The programs were automatically clustered into a hierarchical cluster tree at an

earlier step so that the determination can be made rapidly for each subscriber. The hierarchical cluster tree serves as a decision tree for determining which programs' target profiles are most similar to search profile  $p_k$ : the search for relevant programs begins at the top of the tree, and at each level of the tree the branch or branches are selected which have cluster profiles closest to  $p_k$ . This process is recursively executed until the leaves of the tree are reached, identifying individual programs of interest to the subscriber.

A variation on this process exploits the fact that many subscribers have similar interests. Rather than carry out steps 5-9 of the above process separately for each target profile interest summary of each subscriber, it is possible to achieve added efficiency by carrying out these steps only once for each group of similar target profile interest summaries, thereby satisfying many subscribers' needs at once. In this variation, the system begins by non-hierarchically clustering all the target profile interest summaries of a large number of subscribers. For each cluster  $k$  of target profile interest summary, with cluster profile  $p_k$ , it locates programs with target profiles similar to  $p_k$ . Each located program is then identified as of interest to each subscriber who has a target profile interest summary represented in cluster  $k$  of target profile interest summaries.

#### **Present List of Programs to Subscriber**

Once the profile correlation step is completed for a selected subscriber or group of subscribers, at step 704 the program selection software stores a list of the identified programs for presentation to the subscriber. At a subscriber's request, the processor 406 retrieves the generated list of relevant programs and presents this list of titles of the selected programs to the subscriber, who can then select at step 705 any program for viewing. The list of program titles is sorted according to the degree of similarity of the program's target profile to the most similar subscriber target profile interest summary.

#### **Monitor Which Programs Are Viewed**

The processor 406, at step 707, monitors which programs the subscriber views, keeping track of how much time is spent viewing the program. This information can be combined to measure the depth of the subscriber's interest in the program, yielding a passive relevance feedback score, as described earlier.

The exact details depend on the length and nature of the programs being searched. The computed measure of program attractiveness can then be used as a weighting function to adjust the subscriber's target profile interest summary to thereby more accurately reflect the subscriber's dynamically changing interests.

## 5 Update Subscriber Profiles

Updating of a subscriber's target profile interest summary can be done at step 708 using the method described in copending U.S. Patent Application Serial No. 08/346,425. When an program is viewed, the distribution system 100 shifts the target profile interest summary slightly in the direction of the target profiles of those  
 10 nearby programs for which the computed measure of program attractiveness was high. Given a target profile interest summary with attributes  $u_{ik}$  from a subscriber's target profile interest summary set, and a set of J programs available with attributes  $d_{jk}$  (assumed correct for now), where I indexes subscribers, j indexes programs, and k indexes attributes, subscriber I would be predicted to pick a set of  
 15 P distinct programs to minimize the sum of  $d(u_i, d_j)$  over the chosen programs j. The subscriber's desired attributes  $u_{ik}$  and an program's attributes  $d_{jk}$  would be some form of word frequencies such as TF/IDF, such as from program descriptions or the closed caption text of the program, and potentially other attributes such as the source, and length of the program, while  $d(u_i, d_j)$  is the distance between these  
 20 two attribute vectors (profiles) using the similarity measure described above. If the subscriber picks a different set of P programs than was predicted, the processor 406 should try to adjust u and/or d to more accurately predict the programs the subscriber selected. In particular,  $u_i$  and/or  $d_j$  should be shifted to increase their similarity if the subscriber was predicted not to select program j but did select it, and  
 25 perhaps also to decrease their similarity if the subscriber was predicted to select program j but did not. A preferred method is to shift u for each wrong prediction that the subscriber will not select program j, using the formula:  $u_{ik}' = u_{ik} - e(u_{ik} d_{jk})$ .

Here  $u_i$  is chosen to be the target profile interest summary from subscriber I's target profile interest summary set that is closest to target profile. If e is  
 30 positive, this adjustment increases the match between subscriber I's target profile interest summary set and the target profiles of the programs subscriber I actually selects, by making  $u_i$  closer to  $d_j$  for the case where the algorithm failed to predict

an program that the viewer selected. The size of  $e$  determines how many example programs one must see to change the target profile interest summary substantially. If  $e$  is too large, the algorithm becomes unstable, but for sufficiently small  $e$ , it drives  $u$  to its correct value. In general,  $e$  should be proportional to the measure of program attractiveness; for example, it should be relatively high if subscriber  $l$  spends a long time reading program  $j$ . One could in theory also use the above formula to decrease the match in the case where the algorithm predicted an program that the subscriber did not read, by making  $e$  negative in that case. However, there is no guarantee that  $u$  will move in the correct direction in that case.

10 One can also shift the attribute weights  $w_l$  of subscriber  $l$  by using a similar algorithm:  $w_{ik}' = (w_{ik} - e|u_{ik} - d_{jk}|) / \sum_k (w_{ik} - e|u_{ik} - d_{jk}|)$ .

This is particularly important if one is combining word frequencies with other attributes. As before, this increases the match if  $e$  is positive -- for the case where the algorithm failed to predict an program that the subscriber read, this time by decreasing the weights on those characteristics for which the subscriber's target profile  $u_l$  differs from the program's profile  $d_j$ . Again, the size of  $e$  determines how many example programs one must see to replace what was originally believed. Unlike the procedure for adjusting  $u$ , one also make use of the fact that the above algorithm decreases the match if  $e$  is negative -- for the case where the algorithm predicted an program that the subscriber did not read. The denominator of the expression prevents weights from shrinking to zero over time by renormalizing the modified weights  $w_l'$  so that they sum to one. Both  $u$  and  $w$  can be adjusted for each program accessed. When  $e$  is small, as it should be, there is no conflict between the two parts of the algorithm. The selected subscriber's target profile interest summary set is updated at step 708.

### **Additional Applications**

The above noted application of the data distribution system of the present invention to a cable television system is not intended to limit the scope of the invention. The basic concept disclosed herein is applicable to many applications which entail the use of a subscriber terminal device which has limited data storage capacity and there is limited bandwidth available to transmit data, yet the subscriber terminal device also has access to a plethora of information. A number of such

applications are noted herein to illustrate some significant classes of these further uses of this system.

1. Enhanced Interactive Shopping and Product Demonstration - The present data distribution system allows the use of descriptive information, the principle of locality of reference, and interest-based selection to reduce the memory utilized in a subscriber terminal device. Since the key features of the data distribution system include the automated determination of user interests and pre-fetching, a natural application of the data distribution system is as an enhancement of the present home-shopping technologies, with an interactive menu-based catalog perusal. The interest-based selection features allow merchants to generate user customized catalogs based upon past selections. The data distribution system disclosed can be used at many levels, from choosing particular presentations to selecting "channels" where suitable product information is available. In addition, many manufacturers of power tools, appliances and kitchen apparatus provide video demonstrations of the devices being used, which encourage users to use the device effectively, and encourage non-owners to become purchasers. The data distribution system allows descriptions of the devices to be provided to a large population of potential customers who have an interest in the device.

2. A Handheld Travel Guide Assistant - The present data distribution system can be used to download (via the Web) information relevant to travelers, which information is identified by the user or can be automatically identified on a location basis by GPS (or a local RF signal). This includes not only road and travel conditions but also the location of fuel, restaurants, lodging, points of interest and their associated times of operation. Some local retail stores may also wish to customize catalog information about their products, services and special promotions, or they may wish to provide a comprehensive web based catalog with limited geographically specific off-line selections. Thus, the presently described similarity measures may also be used to selectively present the product selections which are most interesting to the user for any given local catalog. This system can provide mapping information in a multi-channel environment. Consider a 5 channel system, which provides a channel of mapping information, and 4 channels of "zoomed" mapping information at a factor of 2X magnification. The profile



information could be used to detect when particular features are likely to attract the viewer's attention and prefetch mapping information based on these characteristics (e.g., bodies of water, buildings, airports, etc.). With multiple channels, it is easy to see that mapping information could be provided in a multiplicity of directions as well as magnifications. In this system, memory is optimized by selectively downloading information details of sites relevant to the user's interest. Many travel advertisements use video clips or images to show the visual features of local interests. The customization of these presentations based on user interest can ensure a more rapid and insightful selection of destinations, benefitting both the travel industry and customer. This scheme can be used with either a channel selection, indexing or customization service. Furthermore, museums tend to be highly visual, and yet have descriptive information or abstracts for the exhibits for visitors seeking more information about topics and exhibits. The data distribution system disclosed in this application can allow customized museum tours, can develop menus of appropriate selections from a multi-channel environment, or can aid with real-time "channel-surfing" of channels based on user interest.

3. The data distribution system can decide which selections to download and, of the downloaded selections, also decide what granularity of detail is recommended based on the system predicted degree of user interest. Thus, the data distribution system can either download: a) The full video, b) A still graphic, c) A textual description, d) A summary of the textual description, e) The selection title from a directory or menu, f) Do not download the item at all.

4. Real-Estate Previews - Real estate is one of the biggest purchases made by most consumers in their lifetimes, and yet the information delivery system is primitive in the extreme. In particular, single black and white photographs with simple textual descriptions are used to attract customers to a real estate agency or private seller, who then provides an expanded data sheet and maybe a visit to the property if there is interest. This is time-consuming, and when the house is owner-occupied, potentially inconvenient for potential buyers, the seller and the realtor. The data distribution system allows improved selection and presentation of information to potential buyers, including potentially video clips and an increased library of photographs. Depending on the perspective of buyer, realtor or seller, the

data distribution system more effectively pre-screens houses, reduces the cost of selling and time devoted to seeing inappropriate houses, and saves the seller a potentially large amount of inconvenience.

5           5. Smart Card Promotional Discounts - As with today's handheld pager technology, it is possible to build simple receive-only devices at lower cost which embody the concepts disclosed here. In particular, it is a simple technical matter to provide (see "A Distributed Location System for the Active Office," A. Harter and A. Hopper, IEEE Network, Jan./Feb. 1994, pp. 62-70) a low-bandwidth beacon indicating identity, and to couple this with a receiver capable of receiving at higher  
10 rates from a broadcast source, e.g., satellite, broadcast antenna or base station, or the like. If viewed as one aspect of future "smart card" systems, the present invention would allow the downloading of a variety of information useful to the user into such a "smart card" device. In particular, coupon-like electronic credits could be loaded into a smart card to be drawn upon when purchases are made. In this  
15 way, the functions of direct-mail advertising could be replicated in the electronic broadcast facility of the present invention. Purchasing records maintained by the "smart card" device form part of the user profile for this application, suggesting choices of appropriate electronic coupons to be downloaded from the broadcast device. The memory-reduction technology of the present invention has significant  
20 cost and performance advantages for this environment.

### **Summary**

The data distribution system automatically constructs both a target profile for each target object (program) that is broadcast, as well as a "target profile interest summary" for each subscriber, which target profile interest summary describes the  
25 subscriber's interest level in various types of target objects. The system then evaluates the target profiles against the subscribers' target profile interest summaries to generate a subscriber-customized rank ordered listing of target objects most likely to be of interest to each subscriber, so that the subscriber can select from among these potentially relevant target objects, which were  
30 automatically selected by this system from the plethora of target objects available on the data distribution system. This architecture offers several advantages. First, technology upgrades can be done in the core of the data distribution network,

without requiring access to subscriber terminal equipment. Second, there are significant savings in data storage which results from multiplexing this directory information across populations of subscribers. Third, even an extremely limited amount of memory in the terminal adapter can handle complex information streams  
5 and subscriber interests with complete generality.

**WE CLAIM:**

1. A method for the optimization of communication between a terminal device and at least one data source, both of which are interconnected by a communication medium, which terminal device receives data via said communication medium from said at least one data source, which terminal device  
5 transmits communications to said data source via said communication medium, said apparatus comprising the steps of:

producing a prioritized plurality of information segments from each of a plurality of data items which are available from said at least one data source; and  
providing data to said terminal device, identifying a selected data item by a  
10 first of said prioritized information segments of said selected data item.

2. The method of claim 1 wherein said step of providing data to said terminal device comprises:

transmitting, in response to user input at said terminal device transmitting data to said at least one data source identifying said selected data item, data  
5 representative of a second of said prioritized plurality of information segments of said selected data item to said terminal device.

3. The method of claim 2 wherein said step of providing data to said terminal device further comprises:

transmitting, in response to transmission of said second prioritized information segment of said selected data item to said terminal device, data  
5 representative of at least a third of said prioritized information segments of said selected data item to said terminal device.

4. The method of claim 1 wherein said step of providing data to said terminal device comprises:

transmitting, in response to transmission of data to said data source indicative of a user input at said terminal device which requests the entirety of said

selected data item, data representative of all of said prioritized information segments of said selected data item to said terminal device.

5. The method of claim 1 further comprising the step of:

managing said plurality of data items which are available from said data source, to enable efficient access of said plurality of data items by a user at said terminal device.

6 The method of claim 5 wherein said step of managing said plurality of data items comprises:

producing data representative of each of said plurality of data items, which produced data is selected from the class of data item characterizing information which includes: data item access information, data item summary, data item title, data item keywords.

7. The method of claim 6 wherein said step of managing said plurality of data items further comprises:

transmitting, in response to user input at said terminal device transmitting data to said at least one data source requesting identification of all of said plurality of data items, data item characterizing information of all of said plurality of data items to said terminal device.

8. The method of claim 6 wherein said step of managing said plurality of data items further comprises:

transmitting, in response to user input at said terminal device transmitting data to said at least one data source requesting identification of all of said plurality of data items which satisfy criteria input by said user, data item characterizing information of all of said plurality of data items which satisfy said criteria input by said user to said terminal device.

9. The method of claim 6 wherein said step of managing said plurality of data items further comprises:

transmitting, in response to user input at said terminal device transmitting data to said at least one data source identifying selected ones of said plurality of data items, data item characterizing information of said identified selected ones of said plurality of data items to said terminal device.

10. The method of claim 5 wherein said step of managing said plurality of data items comprises:

generating user interest profile data which is indicative of ones of said data items which are likely to be of interest to a user at said terminal device.

11. The method of claim 10 wherein said step of generating user interest profile data comprises:

calculating, in response to said user accessing ones of said plurality of data items, similarity measures to identify other likely data items of interest to said user.

12. The method of claim 10 wherein said step of managing said plurality of data items further comprises:

searching, in response to said user interest profile data, said prioritized information segments of all of said data items to identifying a selected data item  
5 which most likely corresponds to said user interest profile data.

13. The method of claim 1, wherein a plurality of terminal devices are concurrently connected to said communication medium, further comprising the step of:

scheduling activation of said means for providing data to said terminal device  
5 to sequentially serve said plurality of terminal devices according to a determined priority schedule.

14. The method of claim 13 wherein said step of scheduling comprises:  
deciding what information segment is most likely usefully broadcast to each of said plurality of terminal devices.

15. A method of providing a subscriber with program information regarding a plurality of concurrently broadcast programs in a data distribution system which comprises a multimedia broadcast medium which concurrently carries a plurality of programs, which are made available to a plurality of subscribers, which  
5 are connected to the multimedia broadcast medium via respective terminal adapters which contain a directory memory, comprising the steps of:

storing an entirety of directory information in a memory located in said multimedia broadcast medium;

10 excerpting a subscriber specific subset of directory information from said directory information stored in said memory;

transmitting said excerpted directory information to said terminal adapter memory for storage therein;

enabling a subscriber at said subscriber terminal device to access said excerpted directory information stored in said terminal adapter memory.

16. The method of claim 15 wherein said step of excerpting a subscriber specific subset of directory information comprises:

generating subscriber interest profile data which is indicative of ones of said concurrently broadcast programs which are likely to be of interest to a subscriber  
5 at said subscriber terminal associated with said subscriber.

17. The method of claim 16 wherein said step of generating subscriber interest profile data comprises:

calculating, in response to said subscriber accessing ones of said plurality of broadcast programs, similarity measures to identify other likely broadcast  
5 programs of interest to said subscriber.

18. The method of claim 16 wherein said step of managing said plurality of data items further comprises:

searching, in response to said user interest profile data, said prioritized information segments of all of said data items to identifying a selected data item  
5 which most likely corresponds to said user interest profile data.

19. The method of claim 15 further comprising the step of:  
scheduling activation of said step of transmitting said excerpted directory information to sequentially serve said plurality of subscriber terminals according to a determined priority schedule.

20. The method of claim 18 wherein said step of scheduling comprises:  
deciding what excerpted directory information is most likely usefully broadcast to each of said plurality of subscriber terminals.



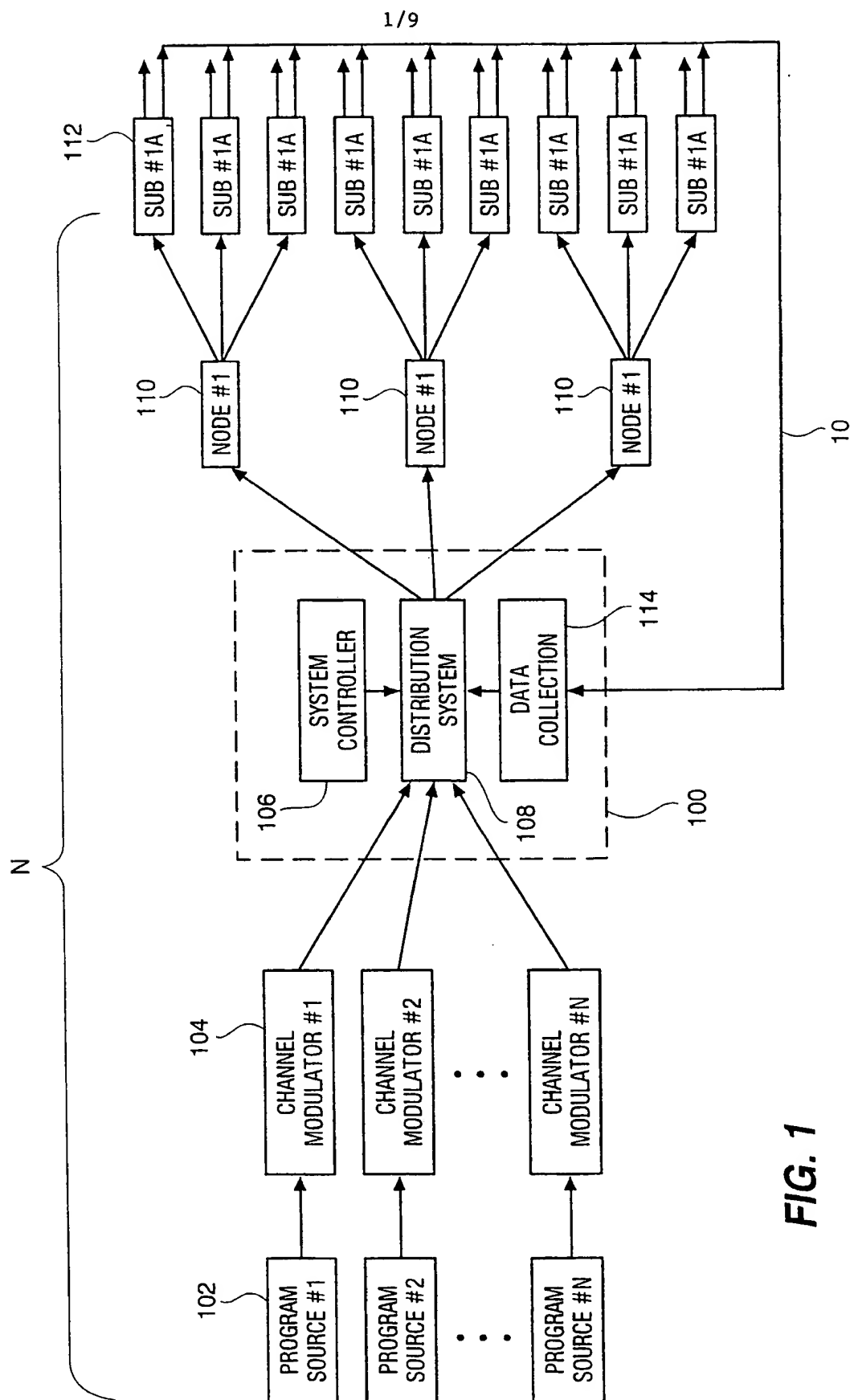
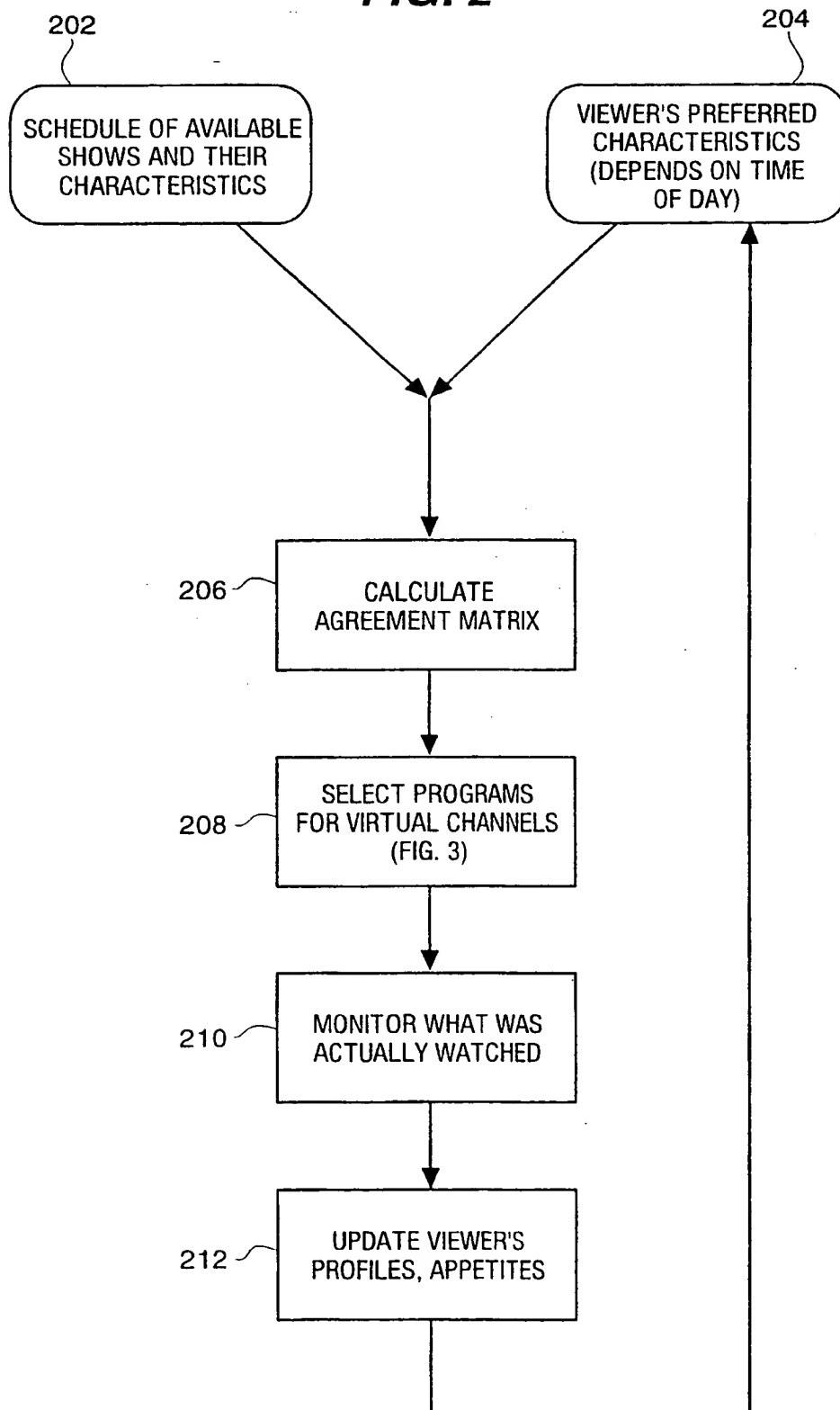
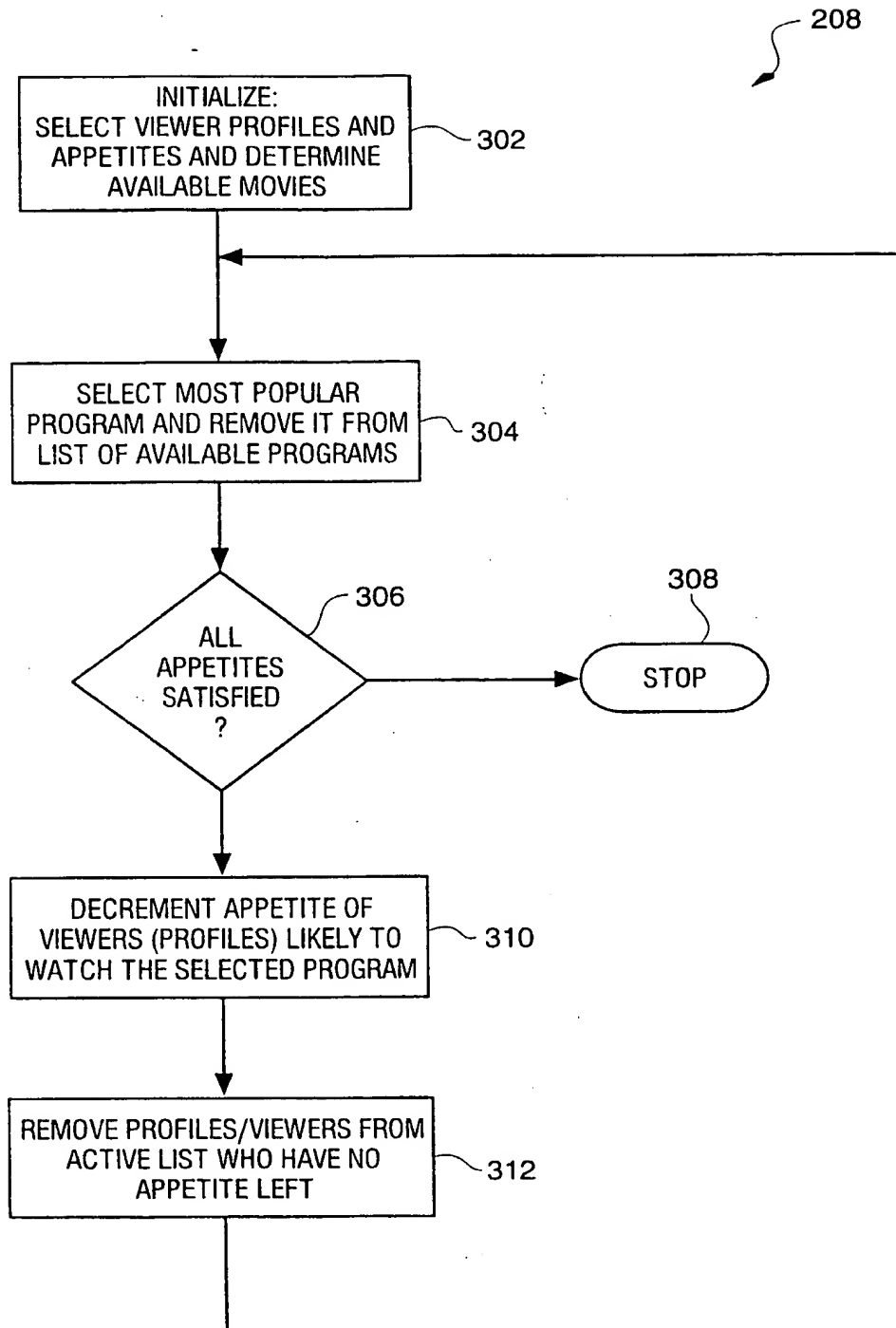


FIG. 1

2/9

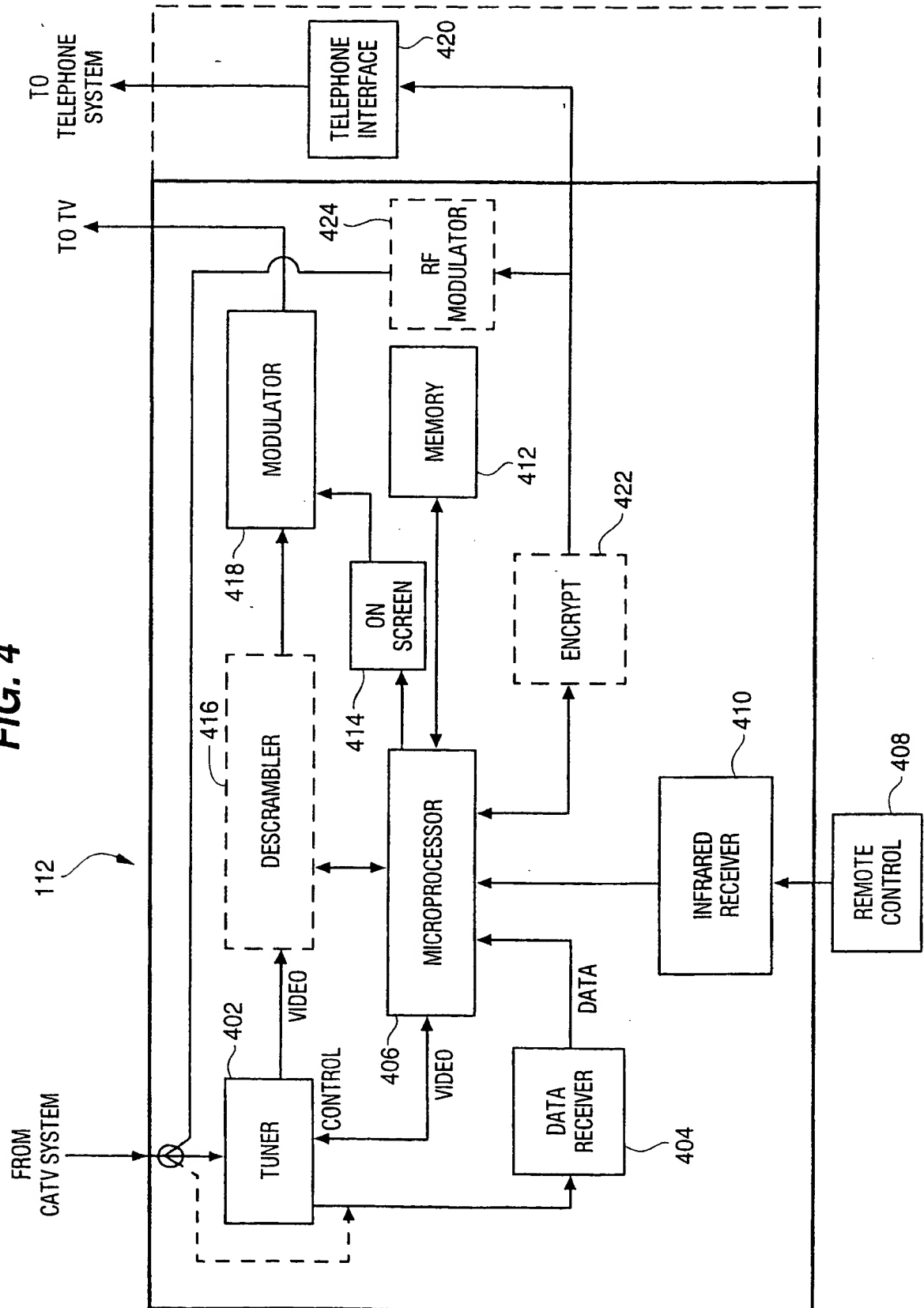
**FIG. 2**

3/9

**FIG. 3**

4/9

FIG. 4



5/9

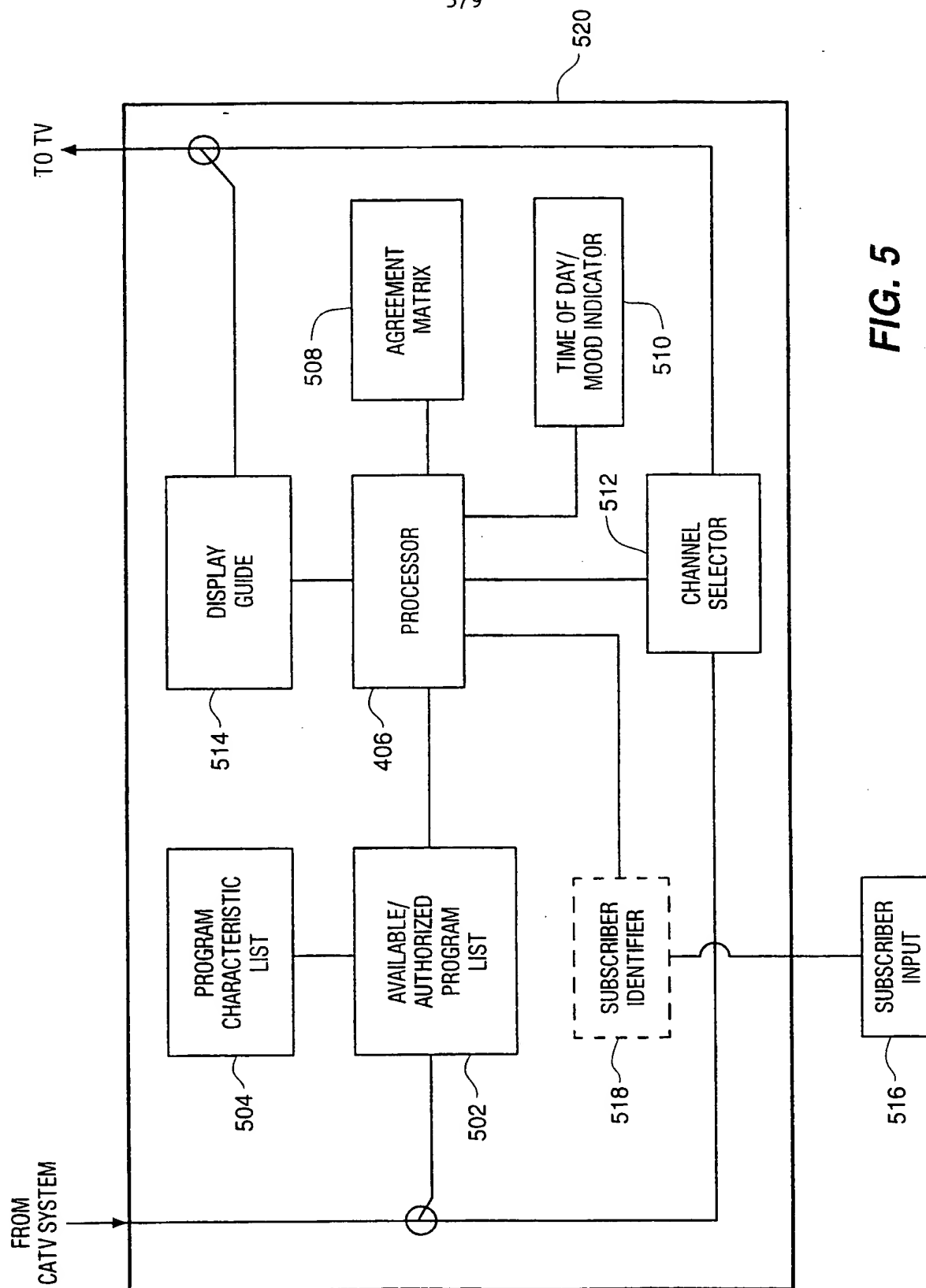
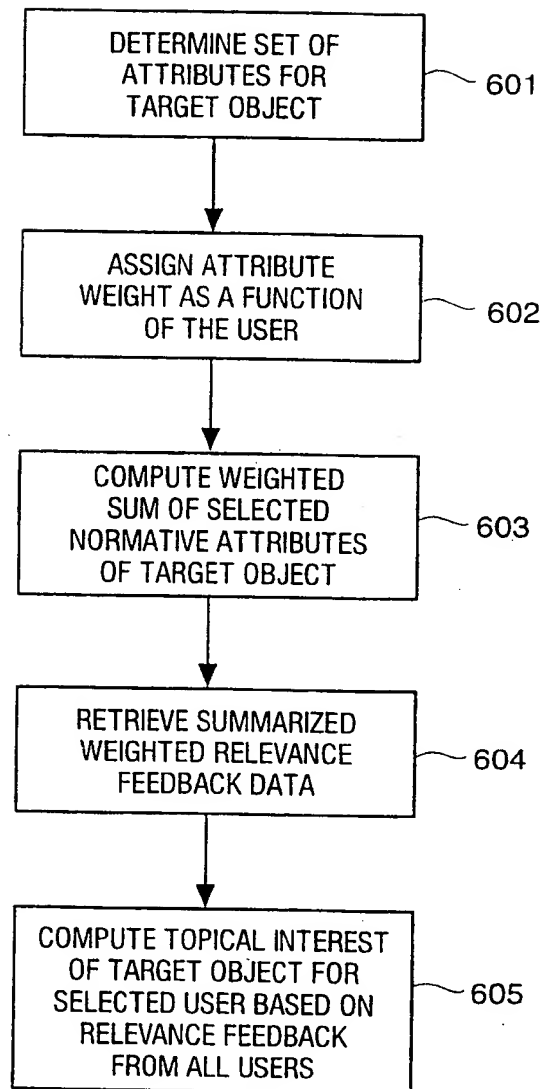
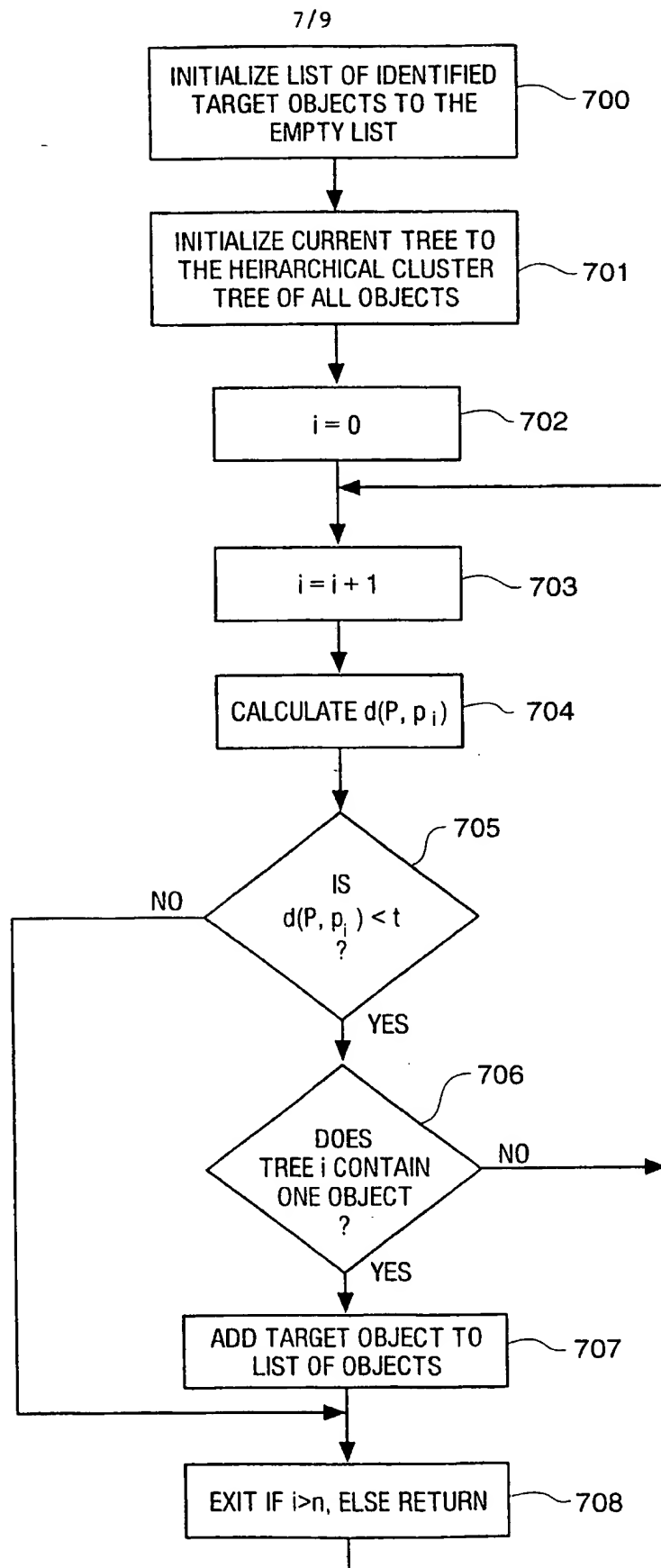


FIG. 5

6/9

**FIG. 6**

**FIG. 7**

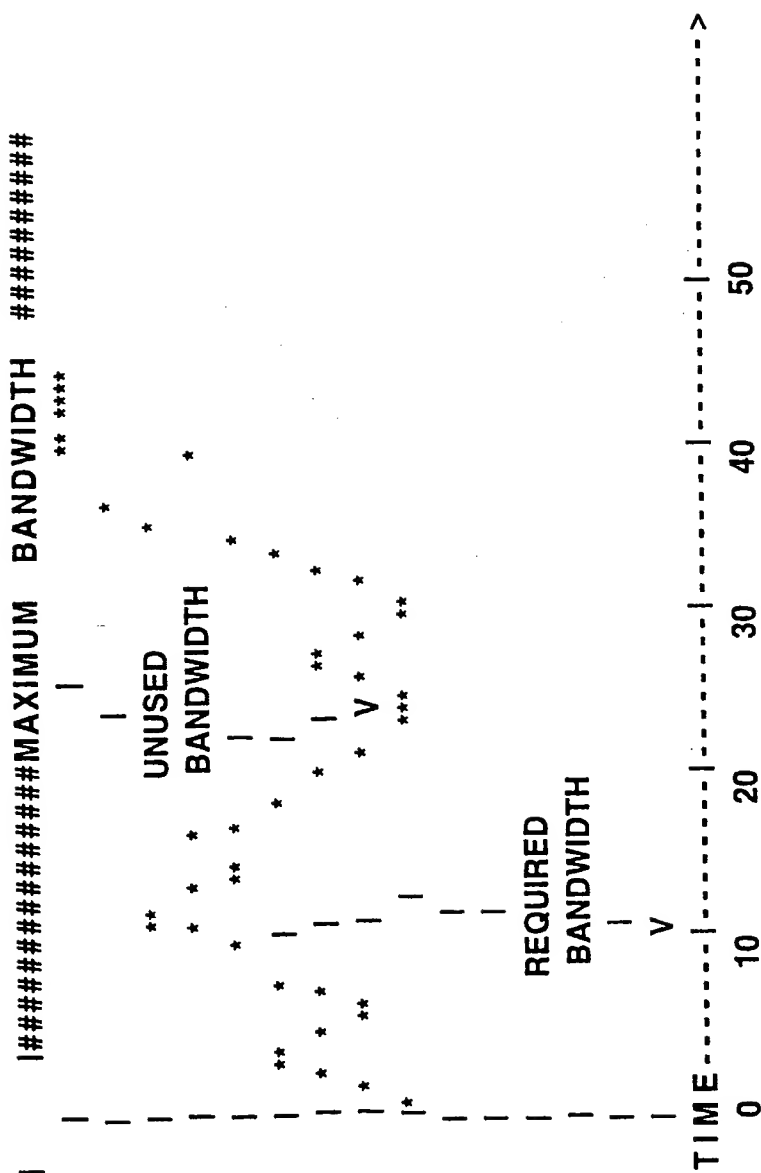
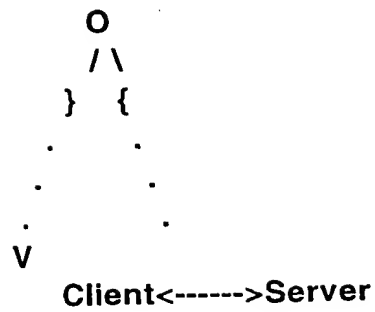
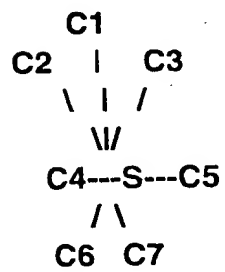


Figure 8



9/9

**Figure 9****Figure10**

# INTERNATIONAL SEARCH REPORT

Internal Application No

PCT/US 98/03181

A. CLASSIFICATION OF SUBJECT MATTER  
IPC 6 H04N7/173

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 H04N H04K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 95 15658 A (DISCOVERY COMMUNICAT INC) 8 June 1995	1,15
A	see page 6, line 26 - page 7, line 17	2-14, 16-20
	see page 15, line 1 - page 16, line 15	
	see page 17, line 19 - page 24, line 8	
	---	
Y	WO 94 10775 A (SCIENTIFIC ATLANTA) 11 May 1994	1,15
A	see the whole document	2-14
	---	
A	EP 0 669 760 A (GRUNDIG EMV) 30 August 1995	15-20
	see column 1, line 7 - column 2, line 46	
	-----	

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

### \* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"Z" document member of the same patent family

Date of the actual completion of the international search

23 April 1998

Date of mailing of the international search report

13.05.98

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Greve, M

# INTERNATIONAL SEARCH REPORT

Information on patent family members

Internat: Application No

PCT/US 98/03181

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9515658 A	08-06-95	US 5600364 A	04-02-97
		AU 1430695 A	19-06-95
		BR 9408211 A	26-08-97
		CA 2177153 A	08-06-95
		EP 0732031 A	18-09-96
		JP 9510327 T	14-10-97
		AU 1333795 A	19-06-95
		CA 2177154 A	08-06-95
		EP 0732027 A	18-09-96
		JP 9506225 T	17-06-97
		WO 9515649 A	08-06-95
-----			
WO 9410775 A	11-05-94	US 5400401 A	21-03-95
		AU 5575894 A	24-05-94
		EP 0682830 A	22-11-95
		JP 8503823 T	23-04-96
		US 5359601 A	25-10-94
		US 5418782 A	23-05-95
-----			
EP 0669760 A	30-08-95	DE 4406091 A	31-08-95
-----			

**This Page Blank (uspto)**